Studies on the Programmable Protein Assembly *via* Genetically Encoded Native Chemical Ligation

A thesis submitted in partial fulfilment of the requirements of the Degree of Doctor of Philosophy at Queen Mary, University of London

By

Wan Ling Wong

School of Biological and Chemical Sciences

Queen Mary University of London

May 2019

Statement of originality

I, Wan Ling Wong, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date: 31st May 2019

Details of publication:

Wright JN, *et al*. (2019) Scalable Geometrically Designed Protein Cages Assembled via Genetically Encoded Split-inteins. *Structure* 27(5):776-784.e4.

Abstract

Programmed protein assembly has vast potential in applications as diverse as bioreactors, smart materials and drug delivery. However, to realise this potential, exact control of the assembly process is required. Thus, this thesis describes generic genetically controlled methods to engineer the self-assembly of geometrically designed protein fusions to form user-defined structures. In particular, it shows how designed fusion proteins can be reacted to form fibres and encapsulations via split-intein mediated native chemical ligation in both one-pot and stepwise syntheses. When compatible fusions are mixed, they react quickly via a split-intein mediated native chemical ligation (NCL) to produce peptide bonded products in high yields (75 % yield). The correctly formed products can be purified to high homogeny and were shown to form the intended user-defined structures.

Acknowledgement

I am heartily thankful to those who have contributed in assisting me towards the completion of my PhD. To begin, I thank my supervisor Dr. Ewan Main for giving me the opportunity to do a PhD with him and his continuous support during my PhD study. His guidance has helped me throughout the research and writing of this thesis. I could not have imagined having a better supervisor and mentor

I am grateful to Dr. John Viles, Dr. Matteo Palma and Dr. James Garnett for their insightful comments and encouragement, as well as their critical questions that have driven me to widen my research various perspectives during progression meetings. In addition, thank you Dr. James Garnett for support in SAXS analysis and Dr. Roberto Buccafusca for mass spectrometry.

My sincere thankfulness also goes to all the members of Main's group. Thank you Dr. Joseph Harvey and Dr. James Wright for teaching me lab skills and Charlotte Frankling (also my gym buddy) for lifting me up when I needed it.

A big thank you to everyone in the 4th floor including those who have left, for all the comfort, encouragement and support, which makes my PhD life tolerable. Special thanks to Ru, Petra, June, Yong Lan, Ambika, Irene, Maddie and Manuela for being such good and supportive friends.

Last but not the least, I thank my family: my parents for believing in me and supporting me financially; my siblings for the spiritual support; and husband, Wai Hong, for absorbing the negativity from me during the hard times.

My PhD would not be as successful without the involvement from anyone of you.

Thank you.

Table of Contents

List of Figures

List of Tables

Nomenclature

2xYT	2 x yeast tryptone media
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
CBD	Chitin binding domain
Cc	Critical concentration
CD	Circular dichroism
CTPR390	Three repeats of consensus tetratricopeptide repeat with binding pocket
CTPRn	Consensus tetratricopeptide Repeat (n = number of repeats)
CV	Column volume
Denat.	Denatured
DTT	Dithiothreitol
E.coli	*Escherichia coli*
FPLC	Fast protein liquid chromatography
FXa	Factor Xa protease
GFP	Green fluorescent protein
Gp	Gp41 DNA helicase inserted split-intein
GST	Glutathione S-transferase tag
GuHCl	Guanidine hydrochloride
H	Histidines-tag
HCl	Hydrochloride
Imp	Inosine-5-monophosphate dehydrogenase inserted split-intein
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kbp	DNA kilo base pairs
KCl	Potassium chloride
kDa	kilo Dalton
LAC	Lactose operon
LB	Luria bertani media
LRR	Leucine rich repeats
M4P	Monofoil-4P
MALDI-TOF	Matrix-assisted laser desorption/ionisation - Time of flight
mAU	Milli-absorbance unit
MESNa	Sodium 2-mercaptoethanesulfonate
$MgCl_2$	Magnesium chloride
$MgSO_4$	Magnesium sulphate
mPa	Milli-Pascal
MS	Mass-spectrometry
MW	Molecular weight
MWCO	Molecular weight cut-off
Mxe Gyr A / MxGA	*Mycobacterium xenopi* DNA Gyrase subunit A
NaCl	Sodium chloride
NADP+	Oxidised NADPH

NADPH	Nicotinamide adenine dinucleotide phosphate hydrogen
NCL	Native chemical ligation
Ni	Nickel
Ni-IDA	Nickel-iminodiacetic acid
NiSO4	Nickel sulphate
OD	Optical density
PBS	Phosphate buffer saline
PBST	PBS supplemented with Tween
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
POI	Protein of interest
rpm	Revolutions per minute
SAXS	Small angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SiRNA	Small interfering RNA
SOC	Super optimal broth with added glucose
TAE	Buffer solution containing a mixture of Tris base, acetic acid and EDTA
TCEP	Tris(2-carboxyethyl)phosphine hydrochloride
TEM	Transmission electron microscope
TEV	Tobacco etch virus
TPR	Tetratricopeptide repeat
Tris	2-amino-2-(hydroxymethyl)-1,3-propanediol
UV	Ultraviolet
Ve	SEC elution volume
Vo	SEC void volume

1 Introduction

1.1 Synthetic biology

There is no stringent definition for synthetic biology. Nonetheless, the simplest and arguably the most precise, is defined by the Synthetic Biology Project, '*Synthetic biology is a) the design and construction of new biological parts, devices and systems and b) the re-design of existing natural biological systems for useful purposes'* ("What Is Synthetic Biology?" 2015). It is also an important method to use in analysing our knowledge of biology *i.e.* do we understand a system to the extent that we can design it according to our needs ("What Is Synthetic Biology?" 2015). In March 2003, the MIT Synthetic Biology Working Group produced a list of potential applications of synthetic biology. These were energy production and storage, new devices and assembly, molecular medical devices, bioreactors, programmable devices and control logic, programmed organisms, smart materials, sensors, complex assembly and terraforming (Community 2003). This PhD fits extremely well into these definitions as it seeks to create a toolkit of designed fusion proteins that are engineered to self-assemble into differing nano-structures. These may have potential applications in bioreactors, smart materials or drug delivery.

1.2 Self-assembly in nature

One method to engineer and design novel self-assembly systems is to take inspiration from nature, as nature uses self-assembled proteins to produce many diverse structures ranging from cages, fibres, network and matrices. Organisms have developed and evolved these useful protein structures to perfection over millions of years. Below are examples of such protein assemblies found in the nature:

1.2.1 Protein cages

Protein cages are hollow protein nanoparticles that have a number of important roles. Perhaps the most well-known cages are the viral capsids, which carry the viral genetic material. They have well-defined, symmetrical, capsule-like structures, with a monodispersed size (Figure 1.1). They are produced from multiple copies of one or a few protein subunits. Due to their function, delivery of their viral genome, these self-assembled capsids need to be robust and well-regulated. Hence, the size of the viral capsid is found to be correlated to the size of the viral genome (Roos et al. 2007; Cadena-Nava et al. 2012).

Two mechanisms have been proposed for the assembly of the capsids (nucleation-and-growth and *en masse*), both involve the RNA polymers assisting the cage assembly (Perlmutter and Hagan 2015). The RNA polymers and cage subunits are produced at the same time via the host's existing organelles. Once the RNA polymers are transcribed and the cage subunits are folded, the RNA interacts with the cage subunits, guiding / recruiting them to surround it. When the cage subunits are in close proximity, they interact with each other to form cages with the RNA polymer in the cavity. The most common capsid organisation is icosahedral. An icosahedron has at least 60 identical subunits forming at least 20 triangular faces. The triangular number (T = x) refers to the number of distinct coat protein subunits present in an icosahedral asymmetric unit. Figure 1.1A is an illustration of icosahedral assemblies of capsid proteins with different triangulation numbers, where the letters denote alternative conformations of the same capsid protein. Figure 1.1B is an example of a self-assembly capsid, the cowpea chlorotic mottle virus (CCMV) (Perlmutter and Hagan 2015). The capsid is made out of three different conformations of protein capsid and assemble to T=3 icosahedral geometry. The diameter of the complete CCMV capsids is 28 nm. This highlights that the assembly of the viral capsid is highly efficient, where only one gene is needed to encode a protein capsid that is able to assemble into a robust protein cage.

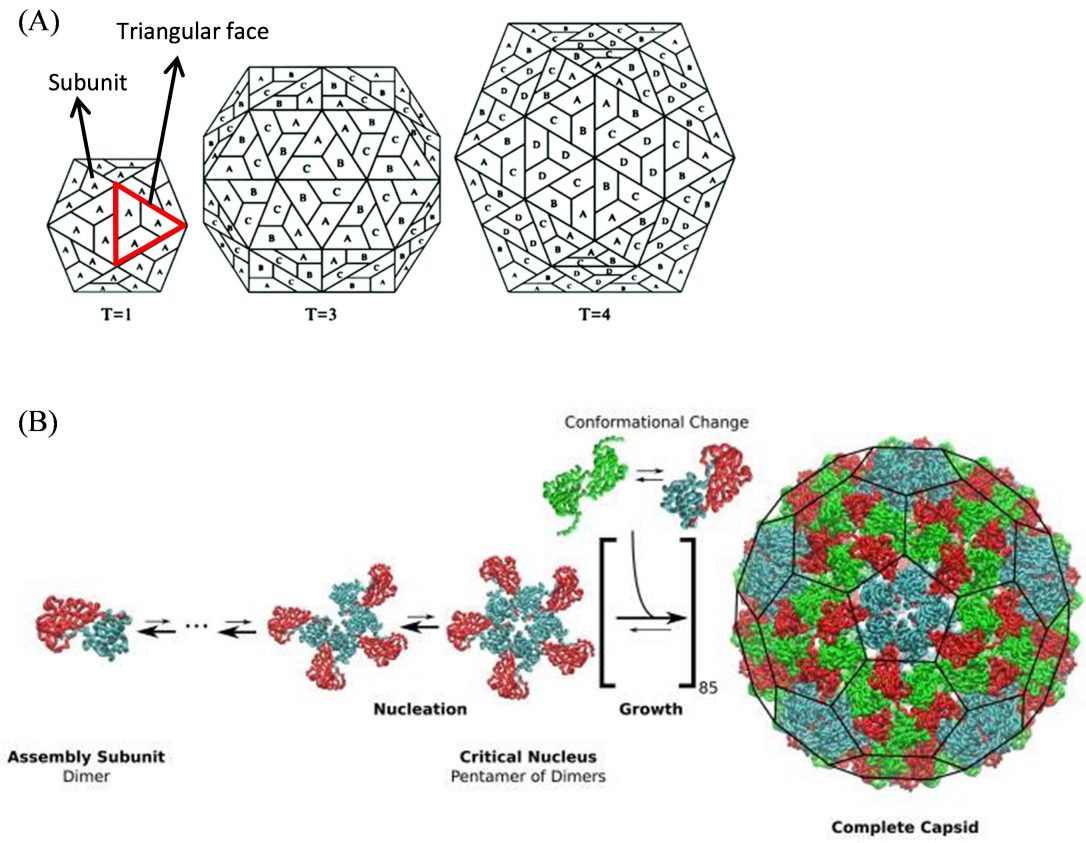


Figure 1.1 Structures and self-assembly of spherical viral capsids. **(A)** Illustration of icosahedral assemblies of capsid proteins with different triangulation numbers, where different letters denote different conformations of the same capsid protein. (Adapted from Mateu, 2013). **(B)** Schematic of the assembly mechanism for cowpea chlorotic mottle virus (CCMV). Different conformations are distinguished by colour (Adapted from Perlmutter and Hagan, 2015).

In contrast to the viral capsid, bacterial microcompartments (BMCs) are a lot larger (typically 100-150 nm in diameter) and are assembled from 10 to 20 different proteins. BMCs play an important role in metabolic pathways: localising the metabolic enzymes in a BMC increases reaction rates, compartmentalising toxic intermediates and enabling tight control through regulated access to the cage. This optimises the metabolic pathway (Brown, Blackwell, and Hammer 2018). The icosahedral BMCs are assembled from shell proteins that come in three main forms, BMC-H, which forms a hexamer, BMC-P, which forms a pentamer and BMC-T, which forms a trimer (Kerfeld and Erbilgin 2015; Parsons et al. 2010; Chiranjit Chowdhury et al. 2014). The combination of BMC-H and BMC-T components tile together to form 20 triangular faces of the icosahedron Figure 1.2. The pores in the BMC-H tiles are smaller, allowing molecules of just a few carbon atoms in and out of the microcompartment (Kerfeld and Erbilgin 2015). Whereas, the BMC-T pores are larger, presumably for movement of bigger molecules, and these can be opened or closed (Kerfeld and Erbilgin 2015). BMC-P is the vertex of the icosahedron Figure 1.2.

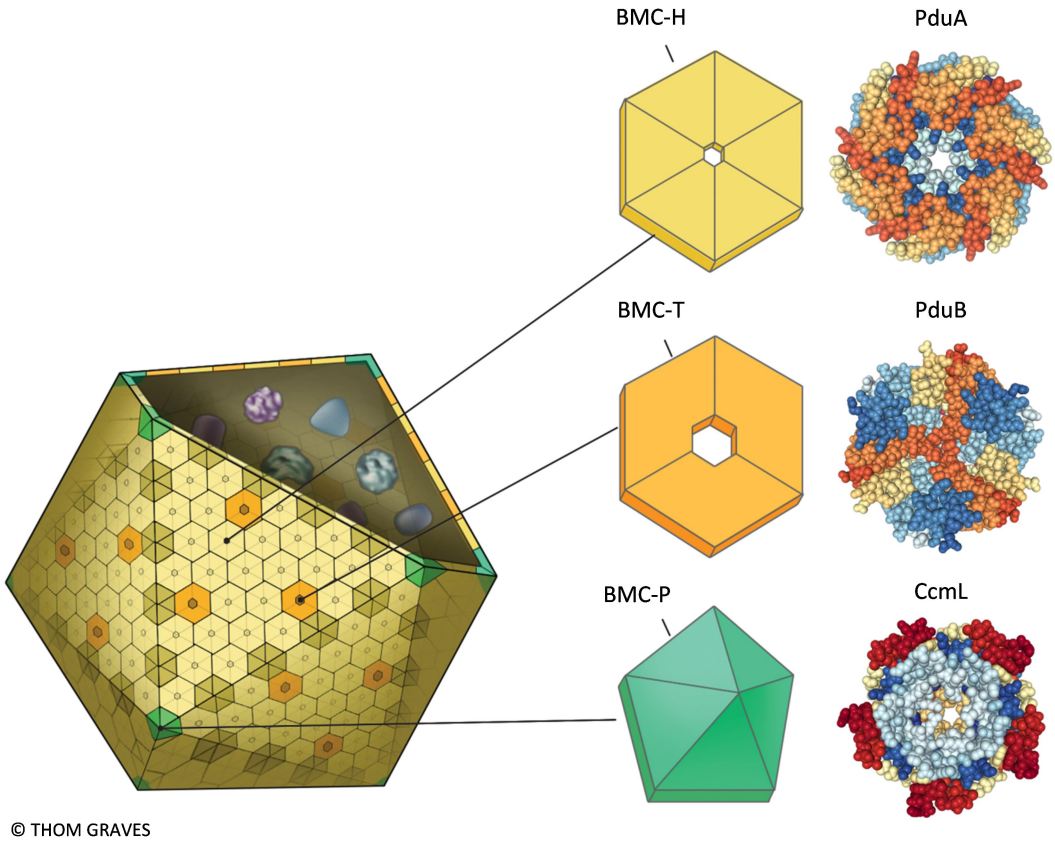


Figure 1.2 Schematic diagram of the icosahedron BMC that assembled from three types of oligomeric proteins, *i.e.* hexamers, trimers and pentamers. Both hexamers and trimers form the faces of the BMC and the pentamers form the vertices of the BMC. PduA (PDB ID 3NGK) is an example of the hexameric BMC-H; PduB (PDB ID 4I61) is an example of the trimeric BMC-T; and CcmL (PDB ID 2QW7) is an example of the pentameric BMC-P.

Another interesting example of protein cages are "Anfinsen cage" chaperonins. Various types of chaperonins are found in prokaryotic cells, endosymbiotic organelles, archaea and eukarya (Spiess et al. 2004). Chaperonins are essential for organismal viability. Anfinsen cage chaperonins provide an enclosed environment to enable either unfolded proteins to fold or misfolded proteins to be unfolded and then refolded (Motojima 2015; Spiess et al. 2004). The most studied chaperonin is the Group I chaperonin, GroEL/GroES found in *E. coli*. Unlike most protein cages, the GroEL/GroES chaperonin is not spherical, it is cylindrical with two distinct chambers and 'lids' (Figure 1.3A). The hollow cylindrical structure with two chambers is assembled from two rings of heptameric GroEL and the lids are made up of heptameric GroES. The two chambers work in tandem, *i.e.* when one is "opened" the other is "closed". In the open GroEL ring conformation, ATP is bound to equatorial domains forcing an expanded tube conformation. This exposes an inner hydrophobic surface of the apical domain (Figure 1.3B) that can bind peptides in an unstructured conformation. Subsequently, GroES interacts with the ATP-mobilised apical domains, causing large apical domain

movements and capping the chamber to form a cage. These movements lead to the release of the unfolded polypeptide into the encapsulated chamber, where re/folding takes place (Figure 1.3C). Finally, ATP undergoes hydrolysis, which weakens the affinity for GroES. Subsequent binding of ATP in the other GroEL ring sends an allosteric signal that ejects GroES, the now folded protein, and the hydrolysed ADP from the chamber (Figure 1.3D-F). At the same time, the opposite ring is now set up to become the folding-active chamber (Figure 1.3E-F).

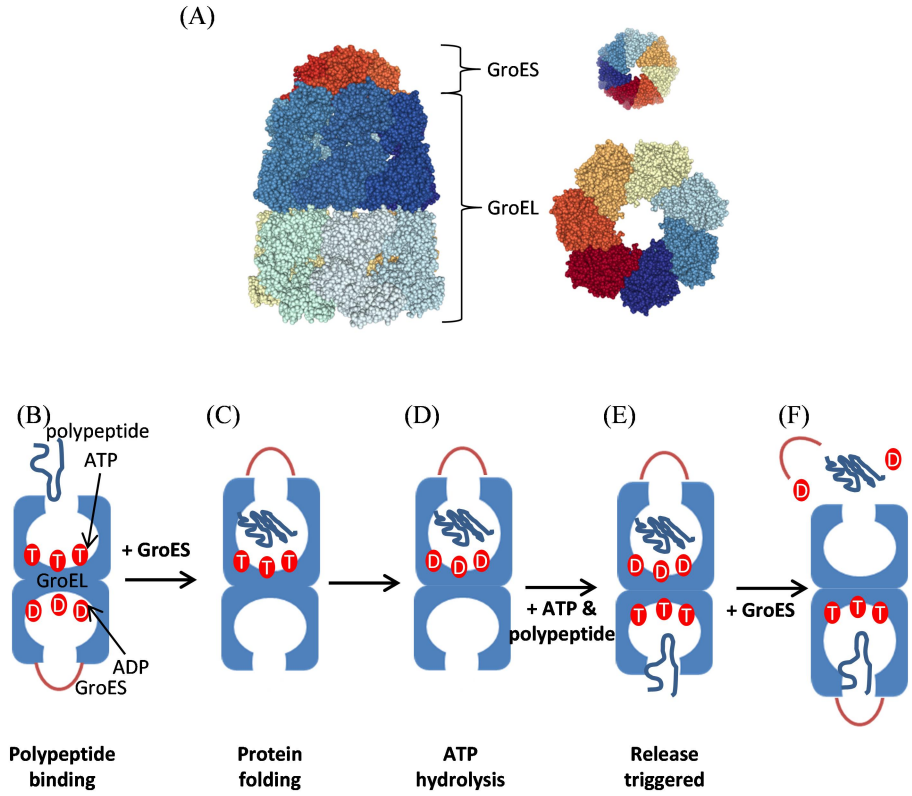


Figure 1.3 **(A)** Crystal structure of chaperonin complex GroEL/GroES (PDB ID 1AON); axial view of GroES (PDB ID 1HX5); and axial view of GroEL (PDB ID 1IOK). **(B-F)** A schematic diagram of the folding of the non-native polypeptide via the GroEL/GroES chaperonin. T-red circle represents ATP; D-red circle represents ADP; blue line represents polypeptide; blue folded line represents folded protein; red curve represents GroES; two blue chambers represent two GroELs. **(A)** The two chambers work in tandem, *i.e.* when one is "opened" the other is "closed". In the open GroEL ring conformation, ATP is bound to the equatorial domains forcing an expanded tube conformation. **(B)** The unstructured polypeptide binds to the inner hydrophobic surface of the apical domain followed by GroES interacting with the ATP-mobilised apical domains, causing large apical domain movements and capping the chamber to form a cage. **(C)** The unfolded polypeptide is released into the encapsulated chamber, where re/folding takes place. **(D)** ATP undergoes hydrolysis, which weakens the affinity for GroES. **(E)** ATP and polypeptide bind to the other GroEL ring. **(F)** GroEL ejectes GroES, the now folded protein, and the hydrolysed ADP from the chamber while the opposite ring is set up to become the folding-active chamber.

A final, simpler, example of intracellular protein cages are Ferritins. Ferritins regulate intracellular metal concentrations to prevent any cytotoxic accumulation and aggregation of metal particles in the cell (Lawson et al. 1991; J.-L. Lee, Park, and Kim 2007). Ferritin is made from 24 monomeric proteins where each subunit consists of a four α-helix bundle

with an extra shorter fifth helix at the C-terminus, Figure 1.4A. Each monomer interacts with 6 adjacent subunits through 3 different interaction faces (forming a dimer, a trimer and a tetramer at different interfaces) to form an octahedral cage with an internal cavity ~ 80 Å in diameter, Figure 1.4B-D (Zhang and Orner 2011). It is thought that the non-polar tetramerisation interfaces acts as an electron transfer channel that can convert insoluble Fe(III) to soluble Fe(II). Once converted, the soluble Fe(II) can exit the ferritin cage via the hollow channel in the trimerisation interfaces. This channel is created by the polar amino acids Asp and Glu (Zhang and Orner 2011).

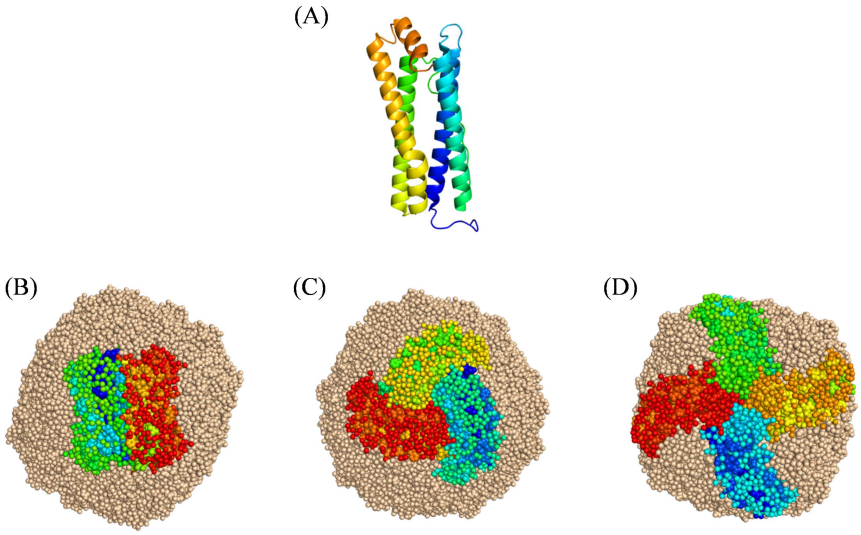


Figure 1.4 Crystal structure of a human ferritin cage (PDB ID 4Y08). **(A)** monomeric ferritin subunit. **(B-D)** Human ferritin cages showing 3 different interfaces, *i.e.* dimer, trimer and tetramer, in rainbow colour.

These examples highlight how nature uses a number of key design features to spontaneously assemble complex ordered cage structures:

(i) Symmetry – In nature, numerous protein domains will form specific dimeric, trimeric and higher polymeric species. These form the vertices and/or sides of a cage. Cages can be assembled from symmetry related homo-oligomeric proteins or from different symmetry related oligomeric proteins. The symmetry enables docking into the cage structure and thus, uses less initial subunits.

(ii) An ordering driving force - No matter the number of protein monomers required to form a cage, they require a specific driving force to dock correctly. In the case of the viral capsids, RNA polymers direct the subunits into position, changing their conformation and enabling protein interfaces to form. In the other examples, the docking interfaces are already present and are driven to associate via key protein-protein interactions.

1.2.2 Fibrous protein assemblies

Unlike protein cages, protein fibres have an elongated shape that can either provide structural support and protection for cells/tissues or more negatively, cause disease. Fibrous proteins are made up of elongated or globular proteins that assemble into fibrous or sheet-like structures. These fibres and sheets are mechanically strong and are water insoluble. There are plenty of fibrous proteins around us. An excellent example is keratin. Keratins can either curl into helices (α-keratins – Figure 1.5A) or bond side-by-side into pleated sheets (β-keratin – Figure 1.5B). The α-keratins are involved in the structure of hair, finger nails, the epidermal layer of the skin, quill, hooves and horns (McKittrick et al. 2012; Bragulla and Homberger 2009; Wang et al. 2016). The β-keratin (also known as fibroins and corneous β-proteins) is much tougher than α-keratin, and produces materials such as silk, claws, scales, feather, beaks and spider webs (Calvaresi, Eckhart, and Alibardi 2016; McKittrick et al. 2012; Bragulla and Homberger 2009; Wang et al. 2016).

Each α-keratin monomer unit ranges from 40-68 kDa and consists of N-terminal, central domain and C-terminal domains (Figure 1.5A). The N and C-termini are responsible for interacting with other filaments / the protein matrix, while the central domain forms the fibres (Wang et al. 2016; Bragulla and Homberger 2009). The α-keratins are keratins folded into right-handed helices that are stabilised by hydrogen bonds. The α-keratins are rich in cysteine residues which promote the formation of a left-handed coiled-coils via disulphide bonds (45 nm long) (McKittrick et al. 2012; Wang et al. 2016). The long coiled-coils stack end-to-end and stagger side-by-side via disulphide bonds to form protofilaments. Two protofilaments laterally associate into a protofibril and four protofibrils forms the intermediate filament with 7 nm in diameter and can link with various matrix proteins. Figure 1.5A summarises the formation of the intermediate filaments from α-keratins (Wang et al. 2016).

For β-keratin, the monomer unit ranges from 10-22 kDa and folds into β-sheets (~2 × 2.3 nm). These are composed of either parallel or antiparallel chains. The peptide bonds of the polypeptide chain force the sheets to arc slightly with respect to each other, termede pleated (Wang et al. 2016). The pleated β-sheets stack in an end-to-end manner forming a distorted left-handed helical shape. Finally, two pleated sheets superpose in opposite directions to form a 3 nm diameter filament. Figure 1.5B summarises the formation of the β-keratin filament from the pleated β-sheets (Wang et al. 2016).

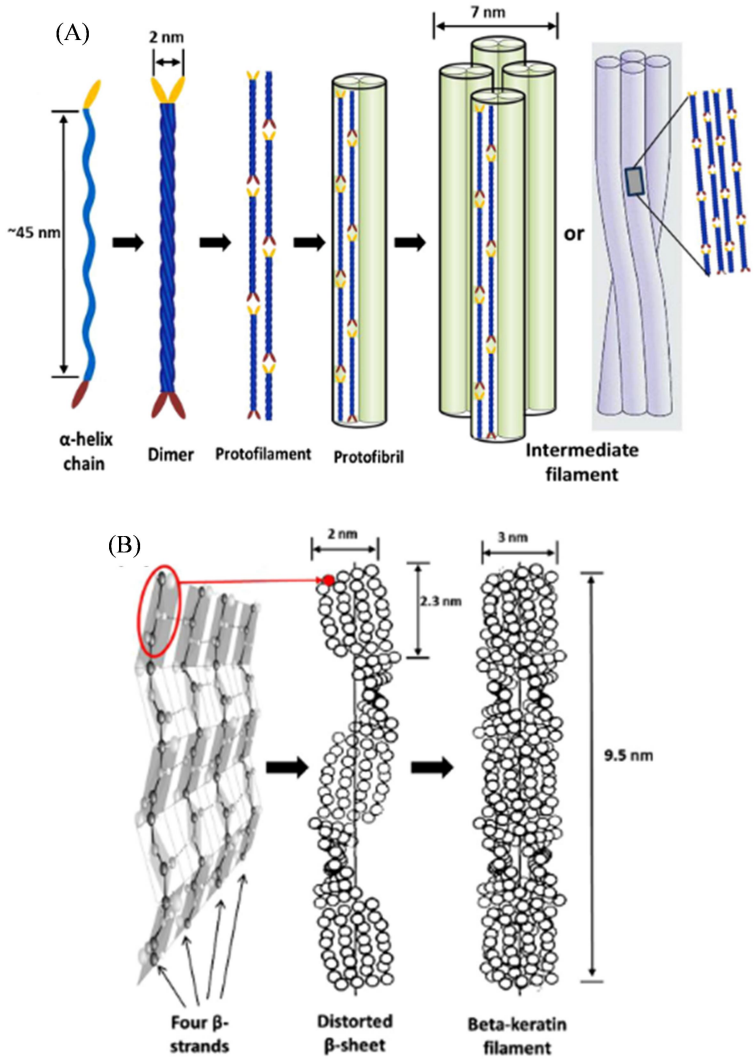


Figure 1.5 Schematic diagram of the formation of the **(A)** α-keratin and **(B)** β-keratin filaments. Image adapted from Wang *et al.*, 2016. **(A)** The α-keratin folds into a right handed α-helix. Two α-helix chains form a dimeric left-handed coiled-coil. The dimers stack end-to-end and side-by-side to form protofilament. Two protofilaments form a protofibril and 4 protofibrils form an intermediate filament. **(B)** β-keratins fold into pleated β-sheets that can stack end-to-end to form a distorted left-handed helical shape. Two distorted pleated sheets superpose in opposite directions to form a 3 nm diameter β-keratin filament.

Post-translation modifications of keratins, such as the formation of disulphide bonds, phosphorylation, glycosylation, inter- and intra-peptide bonds can affect their conformation and therefore the structure of the filaments formed (Bragulla and Homberger 2009). Interestingly, mechanical forces such as stretching, tension and compression can, in certain cases, alter the secondary structure of keratins - from α- to β-conformation (Kreplak et al. 2004).

In contrast to keratin, actin filaments (F-actin) are produced from the globular actin domains (G-actin). The F-actin is a two-stranded helical polymer that is important in defining the shape of the cytoskeleton, cell motility, cell division and muscle contraction. The 41.8 kDa G-actin monomer's atomic structure has been solved (Figure 1.6A) and

shows a globular ATPase with two major domains each of which has two subdomains. The ATP binding side is positioned deep inside a cleft. The ATPase activity of G-actin is very low, but the activity is increased by a factor of 40,000 when it polymerises into F-actin (Kudryashov and Reisler 2013). F-actin assembly starts with a few G-actins forming a nucleus and then polymerising in a head-to-tail conformation via ATP driven docking (Carlier and Pantaloni 1997). Then the F-actin hydrolyses ATP with a rate constant of 0.3 s^{-1}. The release of phosphate from the F-actin initiates intra- and intermolecular conformational rearrangements in the filament that results in the formation of less stable and more flexible ADP-actin filaments (Carlier and Pantaloni 1997). This leads to the depolymerisation of F-actin. The growing end of the filament is known as the barbed end (minus) and the shrinking end is known as the pointed end (plus). Subsequently, a steady state of F-actin is reached. This is known as treadmilling, where a balance between the associations of G-actin to the filament ends is in constant equilibrium with the disassociation of G-actin (Figure 1.6B). Treadmilling occurs when the concentration of the free actin sub-units reaches a point known as the critical concentration Cc, where Cc equals the rate constant for addition of G-actin subunits divided by the rate constant for G-actin subunit loss (Carlier and Pantaloni 1997). This process is spontaneous and slow. However, *in vivo*, there are regulatory proteins to speed up the process. For example, the protein profilin binds to the ATP-bound G-actin, acting as a cap on the barbed end, preventing elongation; the Arp2/3 complex initiates nucleation and introduce branches on existing filaments; and the cofilin proteins can severe filaments into short fragments and promote the disassociation of subunits (Pollard, Blanchoin, and Mullins 2002; Dominguez and Holmes 2011; Carlier and Pantaloni 1997).

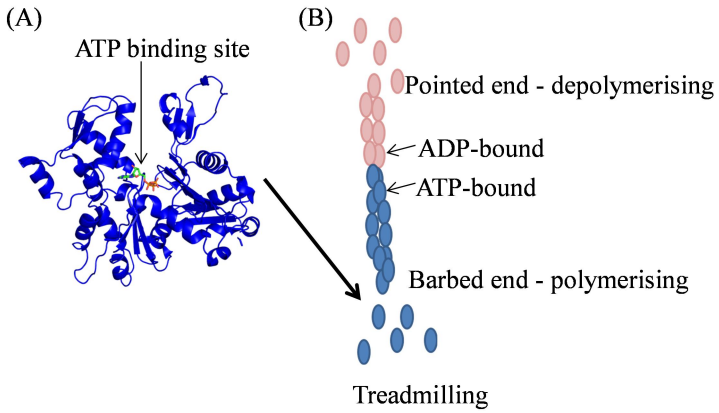


Figure 1.6 The polymerisation of the actin filaments. **(A)** Crystal structure of the G-actin binding an ATP molecule (PDB ID 1ATN). **(B)** A schematic diagram of the treadmilling of an actin filament. Red ovals represent ADP-bound G-actins; and the blue ovals represent ATP-bound G-actins. The barbed end is the polymerisation of the ATP-bound G-actins polymerases forming F-actin while the pointed end is the depolymerisation of the ADP-bound G-actins, shrinking the F-actin. The F-actin hydrolyses ATPs, which results in less stable F-actin, hence depolymerisation.

These examples highlight how nature uses the same design features found in cage assembly to spontaneously assemble fibrous structures:

(i) Symmetry – In general, fibres use only one or two polymerising units. However, the orientation of these defines how fibres extend. Thus, like cages, the protein subunits are symmetry related where the ends of the subunits are orthogonal to ensure head-to-tail polymerisation.

(ii) An ordering driving force – Similar to cage formation, fibre formation requires a specific driving force. The key driving force in fibre assembly, in nature, is driven by non-covalent bond formation. In the case of keratin, post translational modifications play a crucial role in the type of keratin formed, while, the polymerisation of F-actin depends on ATP. Interestingly, many of the driving forces are constitutionally "on" and the regulation of the polymerisation is controlled by the availability of subunits. For example, many regulatory proteins inhibit polymerisation through sequestration.

1.2.3 Protein matrices

Matrix proteins are large molecules tightly bound to form extensive networks via cross-linking insoluble fibrous protein. Elastin is one interesting example of the structural protein that forms an extensive network of elastic fibres, the tropoelastin. It is an important component in tissues, such as ligaments, skin, lungs, tendons and major arteries. The small soluble tropoelastin (50-70 kDa) has highly extensible yet elastic asymmetric coil, with a protruding foot that encompasses the C-terminal cell interaction motif (Figure 1.7A) (Baldock et al. 2011). Although it has been studied extensively, the exact mechanism for assembly has not been completely elucidated. There are, however, several models that have been proposed to explain how elastin produces its unique properties, *i.e.* elasticity, resilience and strength (Anwar 1990; Rauscher 2017; Debelle and Alix 2002). Importantly, all can agree that elastin undergoes alternative splicing during transcription and therefore various tropoelastin isoforms are translated (Green et al. 2014; Debelle and Alix 2002). All of these tropoelastin isoforms are rich in alanine, glycine, valine, proline and lysine residues (Anwar 1990; Debelle and Alix 2002). The lysine residues react to form covalent bonds and thus cross-link tropoelastin monomers together by forming desmosine, isodesmosine and other lysine derivatives (Anwar 1990). It has been found that, between the cross-linked regions, there are mixtures of α-helices, β-strands and

undefined structures (Debelle and Alix 2002). It is believed that these regions contribute to the elasticity. The packing of the tropoelastin is still under investigation because it's hydrated nature is crucial for its elasticity (Rauscher 2017; Debelle and Alix 2002). Figure 1.7B shows the crude model of the packing of tropoelastin into elastin fibres and the changes in structure when it is being stretched.

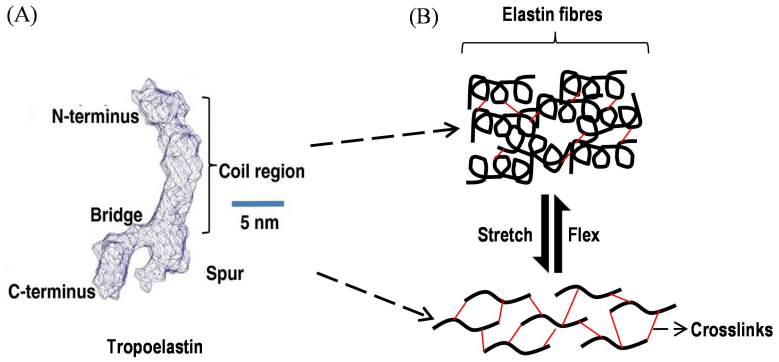


Figure 1.7 **(A)** The structure of tropoelastin solved by SAXS (image adapted from Baldock *et al.*, 2011b). **(B)** A schematic diagram of the packing of elastin fibres in a network and when it undergoes stretching.

Another example of a protein matrix is fibrin, which is important for blood clotting (haemostasis). Fibrin forms a net-like structure capturing plasma, red blood cells, white blood cells and coagulant enzymes *etc*. to stop a wound from bleeding. This is achieved by a 340 kDa glycoprotein complex protein called fibrinogen. Fibrinogen consists of dimers of three proteins, Aα, Bβ and γ. The Aα, Bβ and γ fold into a triple coiled-coil arrangement, where A, B and γ form one end of the triple coiled coil and the α, β and γ form the other. The two halves are held together at the N-termini with disulphide bonds while the C-termini points in opposite directions (Figure 1.8A). During coagulation, thrombin activates fibrin formation by cleaving off small peptides from the N-terminus of the Aα, and Bβ proteins – these are termed small fibrinopeptides A and B (FpA and FpB, respectively) (Figure 1.8B) (Undas and Ariëns 2011; Kattula, Byrnes, and Wolberg 2017; Weisel and Litvinov 2017). This creates "knobs" that insert into "holes" present at the C-terminal of the β and the γ chains on another fibrin. The association of these form protofibrils. Subsequently, the aggregation of protofibrils forms a network. Finally, with the presence of thrombin and calcium ions, transglutaminase factor XIII (FXIII) is activated to cross-link the fibre chains. The FXIII cross-links the fibre chains with isopeptide covalent bonds between lysine and glutamine residues (Weisel and Litvinov 2017; Kattula, Byrnes, and Wolberg 2017). The concentration of thrombin present affects the rate of cleavage of FpB and hence the thickness and tensile strength of the fibrin (Wolberg, Campbell, and Wolberg 2008; Undas and Ariëns 2011). The release of FpB

promotes lateral aggregation and hence increases the thickness of the fibrin formed (Undas and Ariëns 2011).

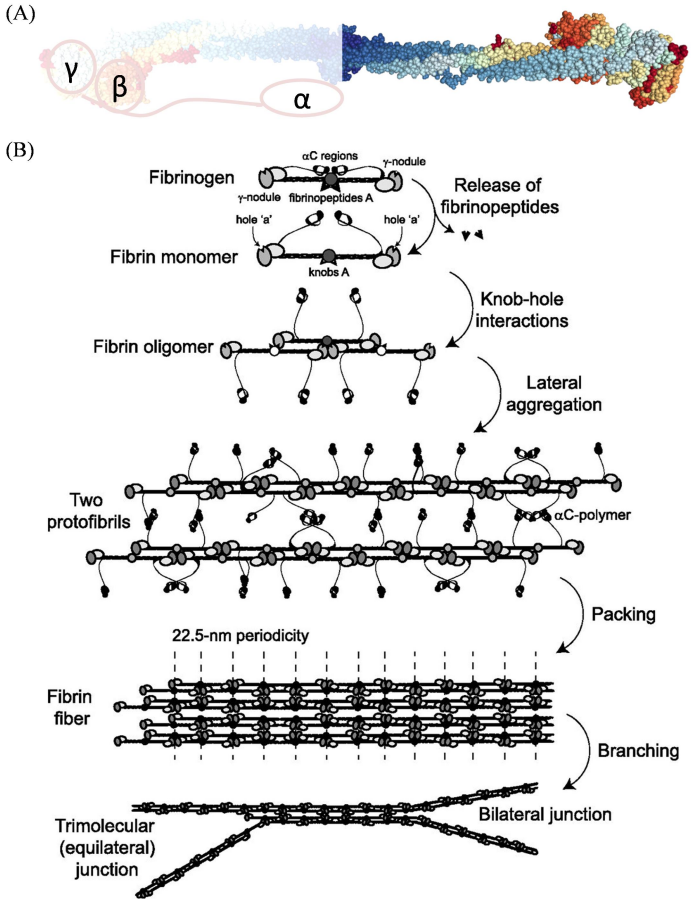


Figure 1.8 **(A)** Crystal structure of fibrinogen without the flexible N-termini of Aα, the FpA and the FpB (PDB ID 3GHG). The added schematics represent the flexible N-termini of Aα, the FpA and the FpB. **(B)** Schematic diagram of the formation of a fibrin fibre (adapted from Weisel and Litvinov, 2017). The thrombin cleaves off the fibrinopeptides allowing the fibrin monomer to form protofibrils *via* knob-hole interactions. Two protofibrils aggregate laterally and pack into fibrin fibres.

Protein matrices again show the same design principles as cages and fibres with some interesting differences:

(i) Symmetry – Similar to fibres, the formation of protein matrices tend to involve one or two crossing-linking protein subunits. These units have symmetry related orthogonal interaction points that order the assembly.

(ii) An ordering driving force – Although, in line with the discussed cages and fibres, a driving force is required for assembly, in these cases both covalent bond formation (elastin) and activation via a protease cleavage event (blood clotting) are used. Covalent bonds were used when a stronger product was required (elastin). Blood clotting requires greater control, thus the subunits are by default inert and "off" until specifically activated.

1.3 Protein building blocks for synthetic biology

Studying the design rules from nature has enabled a number of groups to design and engineer self-assembly systems of protein domains into specific nanostructures. These include: (i) computational docking of protein domains and the design of protein interfaces to produce novel cages and repeat proteins (ii) fusing different oligomeric species together to guide self-assembly and (iii) the re-design of existing modular proteins such as repeat protein motifs to produce fibres and cages. These will be discussed in more detail below.

1.4 Computational docking of protein domains and design of protein interfaces to produce specific nanostructures

A number of recent studies have shown that structures of existing proteins from the protein database can be mined for protein domains that can be computationally redesigned to self-assemble into a user-specified architecture (King et al. 2014; Voet et al. 2014; Hsia et al. 2016; Figueroa et al. 2013; Fallas et al. 2017; Bale et al. 2016). In each case, oligomeris structures deposited in the PDB are docked into the desired nanostructure using RosettaDock. Then, RosettaDesign is used to modify the protein interface regions to increase binding affinity and/or introduce new protein interfaces. The RosettaDesign program was coded by the group of David Baker and has been used for many re-design and protein engineering purposes ("About RosettaCommons" 2015). These include the redesign of individual protein scaffolds into new enzymes and the *de novo* design of individual protein folds (King et al. 2014; Stranges and Kuhlman 2013; Voet et al. 2014; Bale et al. 2016; Fallas et al. 2017).

1.4.1 Nanocages

With regard to novel nanostructure design, one of the most exciting studies was the design of novel nanocages by King *et. al.* (King et al. 2014). Similar to nature, to obtain a caged structure they firstly decided on a symmetry related architecture, *i.e.* tetrahedral, cubic, octahedral, *etc.* Then, all oligomeric protein structures from the PDB were aligned at the vertices of each form. The oligomeric species chosen for each form was such that

the proteins symmetrically contacted each other to form the cage. For example, to create tetrahedral or octahedral cages, 271 trimeric proteins were used and docked with Rosetta's tcdock program (King et al. 2014, 2012). From this, 10 (for tetrahedral) or 20 (for octahedral) trimeric structures had "good" docking (no steric clashing of any two non-bonding atoms in the structures) and these were then subjected to RosettaDesign and FoldIT. Here, the amino acids at the interfaces were mutated *in silico* to generate new interfaces. The end result were pairs of new amino acid sequences, one for each building block. A total of eight tetrahedral and 33 octahedral designs were selected for experimental characterisation. Out of seven tetrahedral and 17 octahedral that expressed natively, one design of each architecture was crystallographically solved Figure 1.9. Recently, these design principles have been taken further and icosahedral cages with 26-31 nm diameters have been produced using two instead of single protein domain components (Bale et al. 2016).

In addition, Hilvert's and Sundquist's groups brought the cages designed by Baker's group a step further by showing its potential in drug delivery. Hilvert and co-workers mutated 6 residues on the cavity surface to arginine of 13 nm diameter cages (O3-33), creating a positively charged cavity (Figure 1.9C) (Edwardson, Mori, and Hilvert 2018). This allowed negatively charged DNA and RNA duplexes to enter the cavity through the ~3.5 nm pores *in vitro*. They successfully transported siRNA, via the cages, into GFP-expressing HeLa cells to knockdown the GFP-expression (Figure 1.9C) (Edwardson, Mori, and Hilvert 2018). Moreover, Sundquist and co-workers re-designed Baker's 60-subunit nanocage, I3-01, to transport cargo from one cell to another. They engineered a fusion protein named EPN-01 by fusing two short peptide sequences to the N and C-terminus of the I3-01 cage forming protein. The N-terminal peptide was capable of membrane binding and the C-terminal peptide was able to recruit the Endosomal Sorting Complexes Required for Transport machinery (ESCRT) (Votteler et al. 2016). The ESCRT machinery catalyses the final membrane fission step required for release from the cell. The study successfully produced ~100 nm vesicles in human embryonic kidney 293T cells. These vesicles contain multiple protein nanocages that closely match the structure of the I3-01 (Figure 1.9D). The vesicles containing the nanocages was successfully released into the cell culture and entered into HeLa cells (Votteler et al. 2016). They further showed that cargo can be packed in the nanocages and delivered to another cell via this system.

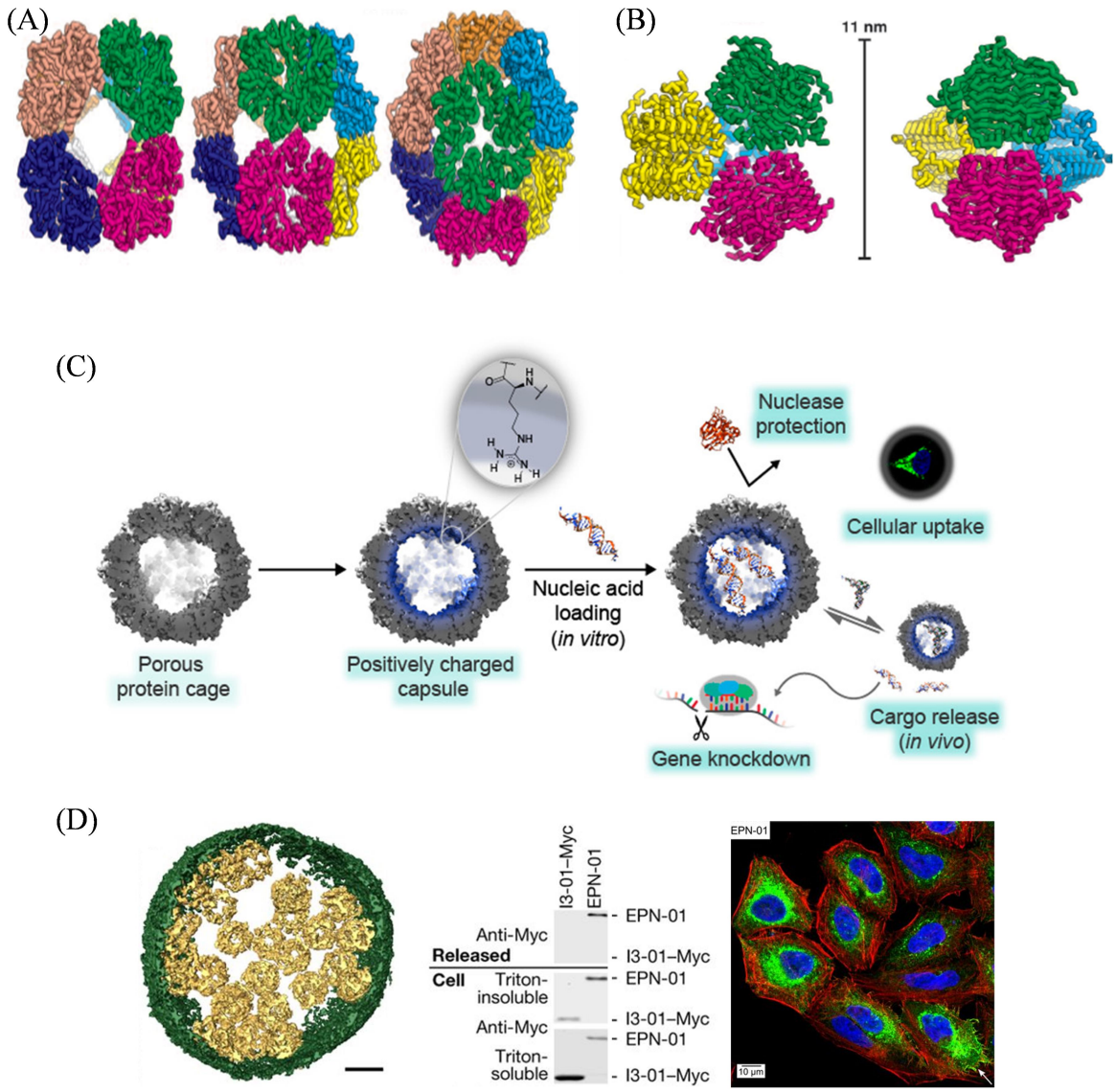


Figure 1.9 Crystal structure of the designed **(A)** octahedral and **(B)** tetrahedral cages (Adapted from King *et al.*, 2014). **(C)** Schematic diagram of modified O3-33 cages designed by Baker and co-workers transporting siRNA into HeLa cells, knocking-down gene expression (Adapted from Edwardson, Mori and Hilvert, 2018). **(D)** On the left, isosurface model of the 3D cryo-EM reconstruction where the EPN membrane is green, individual protein nanocages are gold and the scale bar = 25 nm; in the middle, Western blots showing myc-tagged EPN-01 protein harvested from HEK-293T cell culture supernatants (top blot), cellular EPN-01 proteins in the Triton-insoluble and Triton-soluble fractions (bottom blots) (both I3-01 and EPN were tagged with a myc-tag); on the right, confocal fluorescence images of HeLa cells transfected with myc-EPN-01 stained for myc (green), DNA (blue) and actin (red). Note that EPN-01 is localised primarily in intracellular compartments and at the plasma membrane (white arrow) (Adapted from Votteler *et al.*, 2016).

1.4.2 Repeat protein

Besides nanocages, recent studies have also shown that proteins composed of repeats units (see Section 1.6 for description) can also be redesigned with RosettaDesign. In one study, Baker's group screened for repeat proteins that would dock to form new homo-oligomeric species (C_2-C_6 symmetries). Once suitable candidates were selected, protein-protein interfaces were optimised by RosettaDesign. 96 oligomeric designs were selected for protein expression of which 64 were successfully expressed. Of the 64, 26 were analysed by SAXS with 15 matching the SAXS profile that was expected. Finally of the 15 with matching SAXS data, 5 structures (two dimers, two trimers and a tetramer) were successfully solved to atomic resolution by X-ray crystallography (Fallas et al. 2017). The two dimers and tetramers were built from idealised Ankyrin repeat (ANK) proteins (Figure 1.10A,B and E) while the trimers were built from consensus tetratricopeptide repeat (TPR) proteins and a *de novo* designed repeat protein (Figure 1.10C and D) (Fallas et al. 2017).

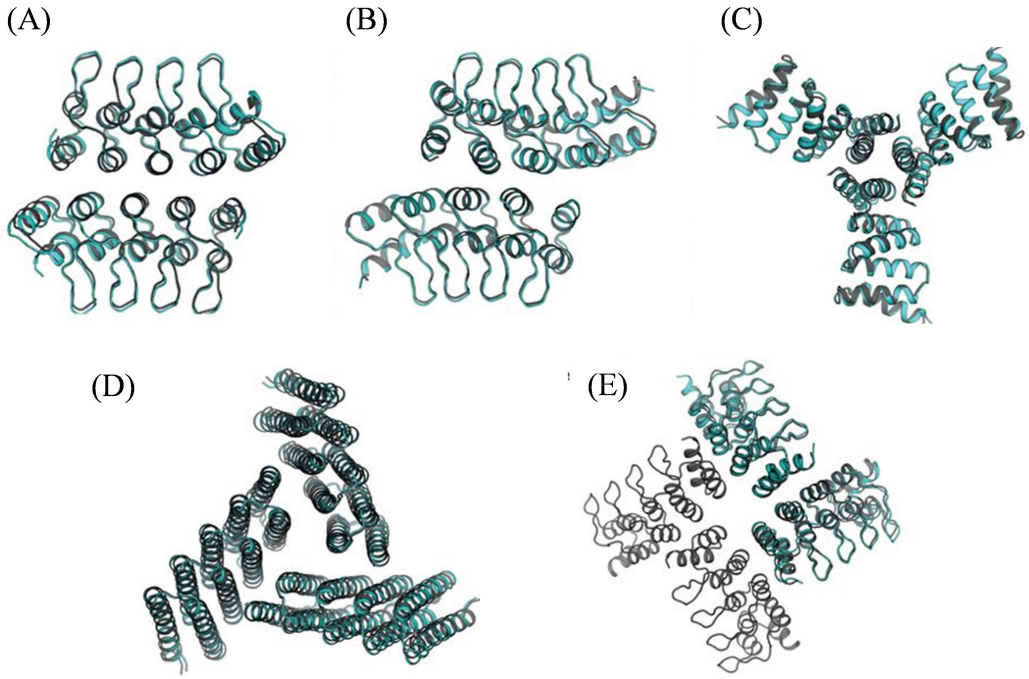


Figure 1.10 Crystal structured of the computationally designed cyclic homo-oligomers based on from **(A, B and E)** Ankyrin repeat proteins, **(C)** consensus tetratricopeptide repeat proteins and **(D)** *de novo* designed repeat proteins.

Unlike Baker's high throughput screening approach, Voet *et. al.* successfully designed and created highly soluble, stable six-fold symmetrical β-propeller protein using a specific template. A natural six-bladed β-propeller *i.e.* the sensor domain of a protein kinase from *Mycobacterium tuberculosis* (PDB code 1RWL, Figure 1.11A), was divided into individual 'blades' for sequence comparison. They used FastML (a webserver for probabilistic reconstruction of ancestral sequence) to identify the possible ancestral blade sequence. The third blade of 1RWL sequence is most likely to be the ancestor and was used as the template to construct a hexameric protein with perfect C_6 symmetry using RosettaDock (RosettaDesign docking simulation software) (Voet et al. 2014). 1000 conformations were created and the one with the best scoring was used to build a single polypeptide chain carrying six identical repeats using the Molecular Operating Environment (MOE). Then, the Rosetta protein modelling suite was used to further analyse the energy of all amino acids at each position before final inspection and tweaking manually. The final symmetrical β-propeller peptide, Pizza6, consists of 42 residues. It is highly soluble and folds stably into a six-fold symmetrical β-propeller protein Figure 1.11B. Multimeric versions of Pizza6 were also created with two and three repeats (Pizza2 and Pizza3), Figure 1.11C and D.

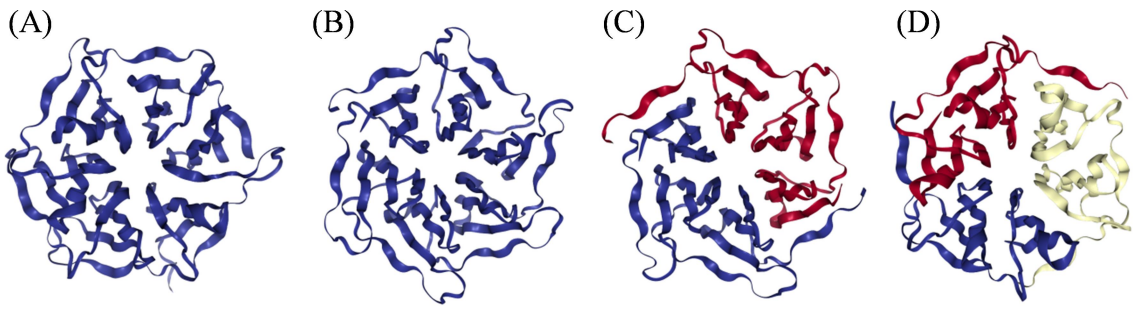


Figure 1.11 Crystal structure of the six bladed β-propeller protein. **(A) T**he non-symmetrical six-bladed 1RWL template protein. **(B)** The symmetrical six bladed protein, Pizza6 (PDB ID 3WW9). **(C)** Dimerisation of the three repeats Pizza3 (PDB ID 3WW8). **(D)** Trimerisation of the two repeats Pizza2 (PDB ID 3WW7).

1.5 Fusing different oligomeric species together to guide self-assembly

In nature, numerous protein domains will form specific dimeric, trimeric and higher polymeric species. These structures spontaneously assemble and are held together by interfaces stabilised by hydrophobic non-covalent interactions, hydrogen bonding and salt bridges. The ability of oligomeric domains to self-assemble makes them interesting candidates for building nanostructures. For example, if one wished to form a cage, specific oligomers can be fused together, causing one oligomer to form the vertices and the other oligomer to form the sides.

1.5.1 Protein fusions of redesigned oligomeric proteins

An excellent example of this type of cage was designed by Lai *et. al.* (Lai, Cascio, and Yeates 2012). They created 16nm cages from the naturally trimeric protein, bromoperoxidase (the vertices), and a naturally dimeric protein, M1 virus matrix protein (the sides). They fused a subunit of the bromoperoxidase and a subunit of the M1 virus matrix protein by an α-helical linker which oriented the symmetry axes of the two components to intersect at an angle half the tetrahedral value of 109.5° (Lai, Cascio, and Yeates 2012). With this angle, twelve designed subunits self-assembled to form a tetrahedral cage as shown in Figure 1.14A. One key point of the design was the orientation of each component; the linker must be rigid enough to keep the components at the correct angle but be flexible enough to allow for assembly.

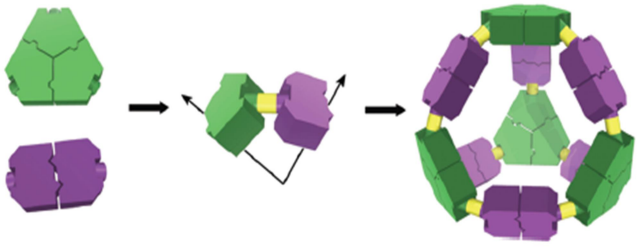


Figure 1.12 Examples of nanostructures formed by oligomers. **(A)** Schematic diagram of self-assembly 16nm tetrahedral cage by fusing a homotrimer and a homodimer by an α-helical linker (adapted from Lai, Cascio, and Yeates 2012).

Another good example of fusing oligomeric domains and other building blocks to build nanostructures is the use of a system called SpyCatcher and SpyTag in conjunction with tetrameric avidin subunits (Fairhead et al. 2014). SpyCatcher and SpyTag are peptides

engineered to form a spontaneous isopeptide bond with each other. They used the Spy system to link chimeric avidin subunits together forming a network. The avidin units used were a mixture of dead streptavidin and traptavidin (an ultra-stable designed biotin binding avidin). They fused SpyTag or SpyCatcher to the C-terminus of the dead streptavidin. The fused dead streptavidin subunits formed a chimeric tetramer with traptavidin subunits in various combinations *i.e.* four dead streptavidin only, three dead streptavidin and one traptavidin, two dead streptavidin and two traptavidin, one dead streptavidin and three traptavidin or four traptavidin only. Different combinations can be separated by ion exchange chromatography. Mixing different chimeric tetramers, desired SpyAvidin i.e. octamers or eicosamers can be assembled (Figure 1.13A) (Fairhead et al. 2014). Different combinations gave rise to different number of traptavidins binding to biotin. Besides that, the SpyCatcher/Tag was used to create multimeric enzyme nanoclusters (Figure 1.13B). Here, the SpyCatcher peptide and its binding partner SpyTag were fused to a dimeric cytochrome P450 monooxygenase mutant (P450BM3m) and a tetrameric glucose dehydrogenase (GDH), respectively (Yin et al. 2019). The fusion proteins successfully self-assembled into a 2D layer of multimeric enzyme nanoclusters that facilitated NADPH regeneration and convert indole into indigo pigment at a much faster rate.

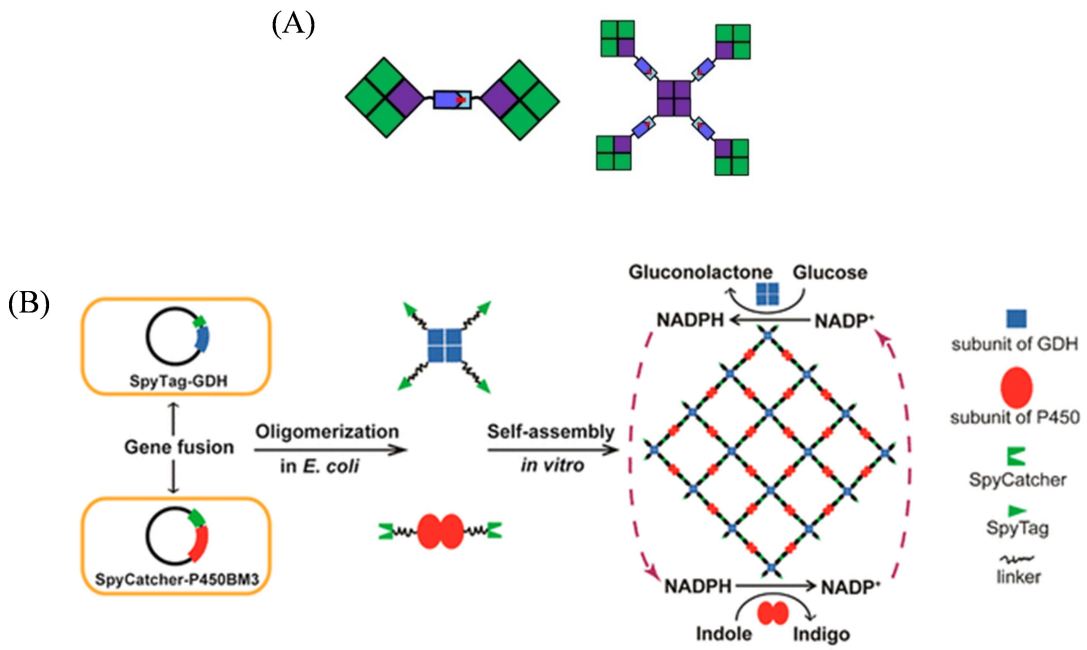


Figure 1.13 Nanostructures formed using SpyCatcher/Tag. **(A)** Schematic diagram of octamer and eicosamer construction. The blue strips represent SpyCatchers; the light blue strips represent SpyTags; the purple boxes represent dead streptavidin subunits; and the green boxes represent traptavidin subunits that bind to biotin (adapted from Fairhead *et al.*, 2014). **(B)** Schematic diagram of the formation of the multimeric enzyme nanoclusters by the protein fusions (adapted from Yin *et al.*, 2019).

Artificial created oligomers can be designed to be used as protein building blocks as well. Kim *et. al.* created different oligomers using green fluorescent protein (GFP) (Kim et al. 2015). They developed a split 'superfolder' GFP system where they split the eleven β-strands of GFP into two *i.e.* 11th β-strand GFP and truncated 1-10th β-strand GFP. They fused the 11th β-strand to the N-terminal of the truncated 1-10th β-strand GFP with an addition of a flexible tri-peptide (Gly-Gly-Thr) linker (Figure 1.14). This linker is designed to provide maximal polymerisation while avoiding intramolecular GFP formation *i.e.* the 11th β-strand GFP will only polymerise with other truncated 1-10th β-strand GFP and form a fluorescently matured GFP as shown in Figure 1.14 (Kim et al. 2015). It can self-assemble to a wide range of oligomers *i.e.* from dimer to decamer (Figure 1.14).

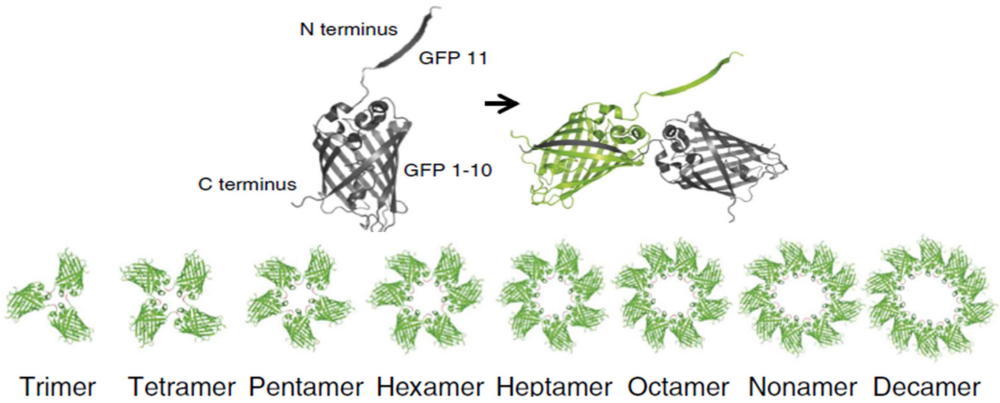


Figure 1.14 Schematic diagrams of the 11th β-strand of GFP fused to the N-terminal of truncated 1-10th β-strands of GFP using a short peptide linker resulting in oligomerisation and differing GFP oligomers. (Adapted from Kim *et al.*, 2015).

1.5.2 Coiled-coils

Besides globular proteins, the elongated coiled-coils have been re-designed to oligomerise into novel nanostructures. Coiled-coils contain a heptad repeat pattern of hydrophobic and charged amino acid residues, forming an alpha-helical secondary structure (Figure 1.15). The heptad repeat pattern is a sequence of seven amino acids in H P P H P P P manner, here H represents hydrophobic residues and P represents hydrophilic residues. The hydrophobic residues force the helices to gently coil around each other in a left-handed direction, forming an amphipathic structure as shown in Figure 1.15 (Armstrong et al. 2009; Harbury et al. 1998). The amino acids at hydrophobic and hydrophilic positions play a role in oligomerisation of coiled-coils. In particular the exact amino acids at positions **a - d** and **e - f** tend to be the main determinants of the oligomeric state. For example, the dimeric leucine zipper encodes for a leucine at the **d** position (Baxevanis and Vinson 1993; Ciani et al. 2010).

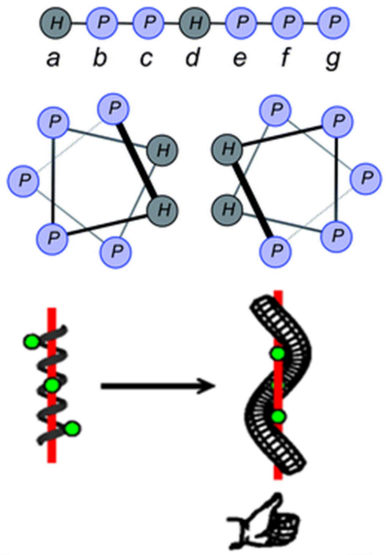


Figure 1.15 Schematic diagram of an heptad repeat coiled-coil. The regular α-helix produces a left-handed coiled. H and green circles represent hydrophobic residues, and P represents hydrophilic residues (Adapted from Harbury *et al.*, 1998; Armstrong *et al.*, 2009)**.**

With regards to nanostructure design, Fletcher *et. al.* successfully designed coiled-coils to self-assemble into cages with a diameter of 100 nm. The system uses two *de novo* coiled-coil building blocks: (i) a homotrimer called CC-Tri3 and (ii) a heterodimer called CC-Di. Both the CC-Tri3 and CC-Di possess cysteines on each of their coiled-coils Figure 1.15. To create the assembling system the CC-Tri3 is incubated with either half of the CC-Di coiled-coil. This creates two complementary trimeric hubs CC-Tri3—CC-Di-A and CC-Tri3—CC-Di-B via disulphide linkage. When mixed together, these two hubs will self-

assemble to form hexagonal networks that closes to form cages (Figure 1.16A**)** (Fletcher et al. 2013). Moreover, the homo-trimer coiled-coil can be decorated with peptides or proteins via fusion at their termini. For example, recombinantly produced fusion proteins of cysteine-free GFP on either terminus of the Cc-Tri3 have successfully formed hubs with CC-Di and assembled into cages when mixed (Ross et al. 2017).

On the other hand, instead of using multimeric assembly, Gradišar *et. al*. used a single polypeptide chain containing 12 coiled-coil segments to fold into a tetrahedral cage (Gradišar et al. 2013). This strategy utilises the ability of the coiled-coil segments to form defined anti-parallel and parallel dimers. The coiled-coil segments were separated by flexible peptide hinges, via a Ser-Gly-Pro-Gly motif. The chain path follows the edges of a tetrahedron, passing through each edge exactly twice, so that the polypeptide chain interlocks the structure into a stable shape formed by the six coiled-coil dimers (Figure 1.16B). The protein folds to form a tetrahedral nanocage structure named TET12. Recently, the group has developed the system to produce a more soluble tetrahedron, a rectangular pyramid (that contains a four-branched vertex), and a triangular prism that can self-assemble *in vivo* and *in vitro* (Ajasja Ljubetič et al. 2017).

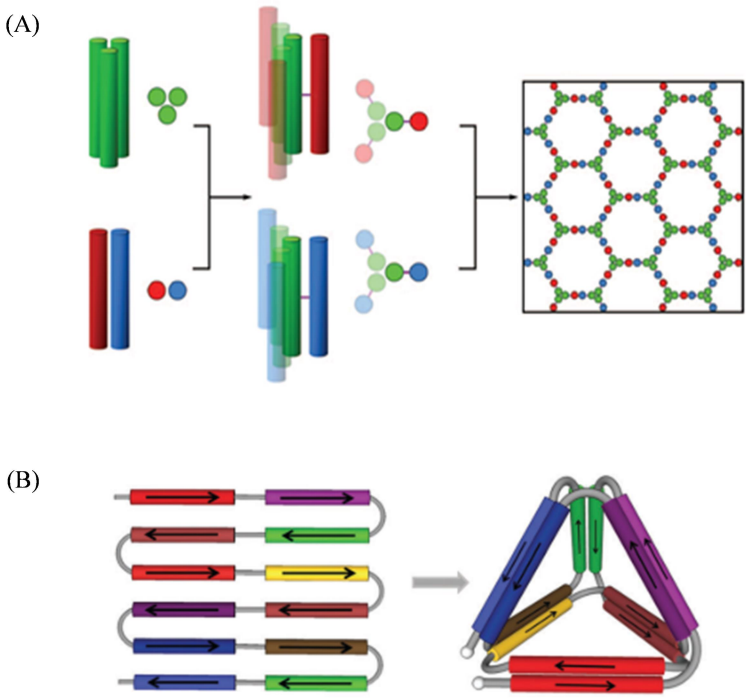


Figure 1.16 **(A)** Schematic diagram of the self-assembly CC-Tri3—CC-Di-A and CC-Tri3—CC-Di-B. The CC-Di-AB is linked to CC-Tri3 via disulphide bonds producing complimentary trimeric hubs. When mixed together, it self-assembles into cages. The green circles represent CC-Tri3, the homotrimer coiled-coil; the red circles represent CC-Di-A, the acidic alpha helix of the heterodimer; and the blue circles represent CC-Di-B, the basic alpha helix of the heterodimer (Adapted from Fletcher *et al.*, 2013)**. (B)** Schematic diagram of the self-assembled TET12. A single chain of 12 coiled-coil segments separated by flexible linker folds into a tetrahedron by forming specific antiparallel and parallel dimers (Adapted from Ljubetič *et al.*, 2016).

1.5.3 Fusion protein of coiled-coils and oligomeric proteins

Several groups have successfully utilised coiled-coils as the actual driving force to drive assembly of structural proteins into differing protein nanostructures (Ross et al. 2019; Sciore et al. 2016). For example, octahedral protein cages with ~18 nm diameter have been formed by fusing tetrameric four-heptad coiled-coil to the C-terminus of the trimeric esterase protein (Figure 1.17A) (Sciore et al. 2016). Here, the esterase proteins act as the sides and the coiled-coils are the vertexes. In contrast, nanotubes with a diameter of approximately 24 nm and various lengths up to 600 nm were assembled when trimeric four-heptad coiled-coil fused to the C-terminus of the pentameric cholera Toxin B (CTB) (Figure 1.17B) (Ross et al. 2019). Interestingly, weak interactions of the CTB pentamers were observed when the coiled-coil brought them into close proximity.

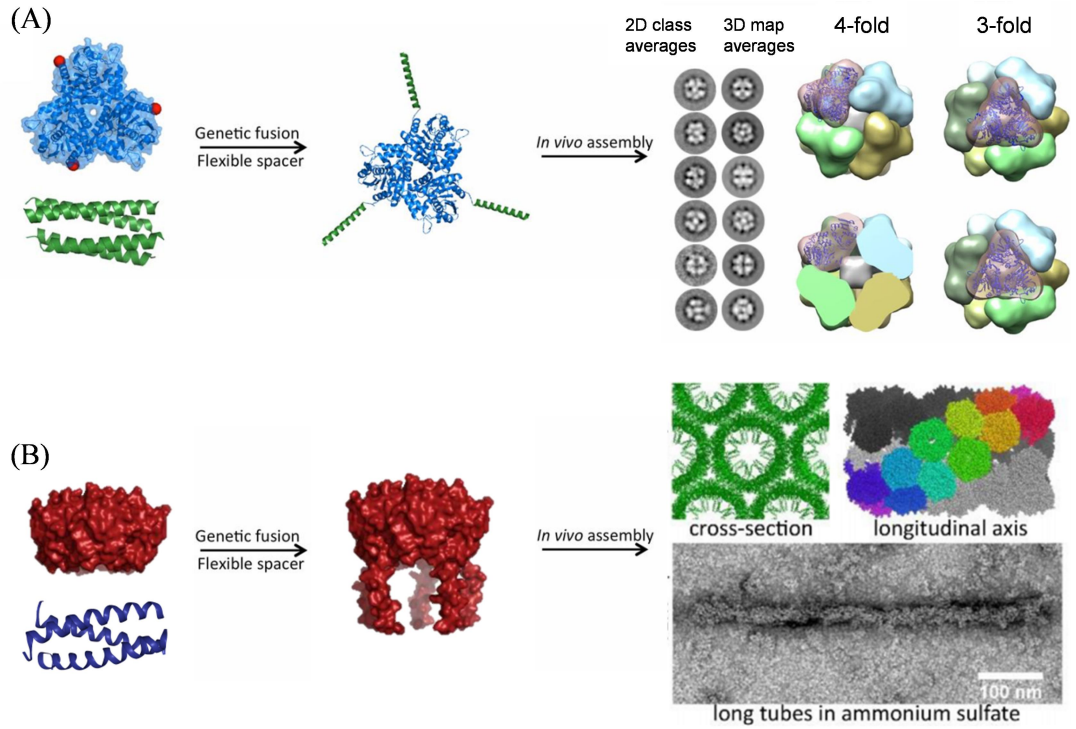


Figure 1.17 Diagram of the self-assembly of **(A)** nanocages and **(B)** microtubes. **(A)** The fusion of homotrimer esterase (PDB 1ZOI), where the C-termini are indicated with red spheres, and homotetramer coiled-coil (PDB 3R4A) self-assembled into nanocages. The nanocages formed were analysed by negative stained EM. **(From right) R**epresentative 2D class-averaged images of the cages and projections generated from the 3D electron density map, and the reconstructed electron density viewed along the fourfold and threefold axes with one esterase trimer shown modelled into the electron density. The lower images show a slice through the electron density (Adapted from Sciore *et al.*, 2016). **(B)** The fusion of homopentamer CTB (PDB 3CHB) and homodimer coiled-coil (to represent homodimer coiled-coil, PDB 1COI was used here) self-assembled into microtubules. The microtubules formed were analysed by crystallography and TEM. **(Top right)** A section of the crystal structure looking down the z-axis, showing packing of the tubes and the inner lining of the coiled-coils. **(Top left)** Longitudinal axis of the crystal structure. **(Bottom)** Incubation of 58 μM CTB-coiled-coil fusions in 0.5 M ammonium sulphate gave rise to tubular structures under TEM with a diameter of approximately 24 nm and various lengths up to 600 nm.

1.6 The re-design of existing modular proteins to produce user defined nanostructures

Repeat proteins are a large family of proteins fold that are composed of repeated units of 20-30 amino acids. These units encode for secondary structural elements that stack together and form the final three dimensional fold. Thus, such protein repeats thus possess very symmetrical structures, with series of regularly spaced secondary structural elements. Interestingly, the more repeats in a folded domain, the greater the proteins stability. Figure 1.18 shows an indicative selection of different repeat proteins: antifreeze protein (AFP), ankyrin repeat (ANK), Armadillo repeat (ARM), Huntingtin, elongation factor 3, protein phosphatase 2A and the yeast kinase TOR 1 (HEAT) repeat, hexapeptide repeat, leucine rich repeat (LRR), tetratricopeptide repeat (TPR) and WD40 repeat (Main, Lowe, et al. 2005).

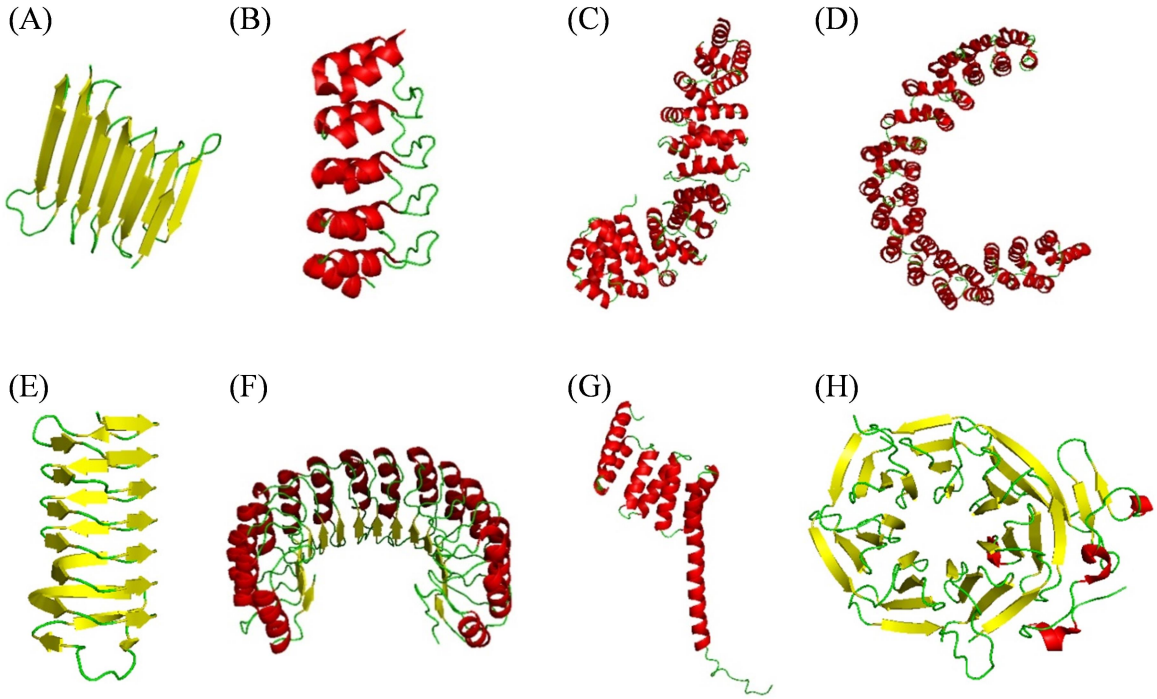


Figure 1.18 Ribbon representations of repeat proteins. **(A)** The crystal structure of Rhagium inquisitor AFP (PDB code 4DT5). **(B)** The crystal structure of an improved thermally stabile designed ANK with a redesigned C-capping module (PDB code 2XEE). **(C)** The ARM repeats subunit from murine β-catenin (PDB code 3BCT). **(D)** The HEAT repeats domain from a protein phosphatase 2A holoenzyme with B55 subunit (PDB code 3DW8). **(E)** The crystal structure of hexapeptide repeats (1M8N). **(F)** The LRR region from a ribonuclease inhibitor (PDB code 1DFJ). **(G)** The crystal structure of a TPR (PDB code 1A17). (H) The crystal structure of the C-terminal WD40 domain of TUP1 (PDB code 1ERJ). All repeat proteins are coloured according to secondary structure where red for α-helices, yellow for β-strands, and green for loops.

There has been great success in the design of repeat proteins by creating proteins of tandemly arrayed consensus repeats. The consensus residues correspond to the most frequent amino acid residue found at each position in the repeated unit. For example, Main *et. al.* used consensus-based design to successfully engineer novel tetratricopeptide repeat proteins (TPRs) (Main et al. 2003). They aligned all known TPR sequences and selected the most frequent at each position. Currently, the TPR motif can be found in over 1556 different proteins with differing number of repeats (D'Andrea and Regan 2003; Finn et al. 2014). Similarly, Mosavi *et. al.* and Kohl *et. al.* designed proteins based on a consensus ANK repeat (Mosavi, Minor, and Peng 2002; Kohl et al. 2003). Interestingly, these designs produce "bare" scaffolds *i.e.* while the design process selects for those amino acids important for structural integrity, it removes all amino acids important for binding functions. However, binding function can be easily engineered into the structures through directed evolution (Schlinkmann and Plückthun 2013; Egloff et al. 2014) or a variation of the above consensus approach (Grove et al. 2010, 2012; Speltz, Nathan, and Regan 2015).

Given their designability, repeat proteins have also been successfully used as building blocks of various "smart" nanostructures. Examples include:

(i) Main's group utilisation of three repeats of CTPR (CTPR3) as a building block and intein meditated Native Chemical Ligation (NCL) to form fibrous nanostructures. Two different methods were experimented, *i.e.* (1) uncontrolled fibres extension that successfully produced various lengths of fibres; and (2) iterative fibres extension where a precise extension of CTPR3 was produced via stepwise addition. Both methods were successful and the details will be discussed further in the next section (Section 1.7).

(ii) The use of an 18 repeat consensus TPR protein (called CTPR18) to form a hydrogel by adding multivalent cognate- polyethylene glycol (PEG) cross-linker (Figure 1.19A) (Grove et al. 2012). Grove *et. al.* created a gene that encoded for 3 repeat TPR units that bound a DESVD peptide (Binders - B) with 3 repeat units that were without binding sites (Spacers - S). The order was B-S-B-S-B-S (Figure 1.19A). The S acts as a spacer so that the cross-linking sites are equally spaced and offset by 120° along the helices (Figure 1.19A). 4-arm star-like PEG-DESVD cross-linkers are created by coupling

peptides through an N-terminal cysteine and DESVD peptide is added at the C-terminus. The gel was formed when PEG-DESVD was added to CTPR18. Interestingly, when 500mM sodium chloride (NaCl) was added, the gel dissolved as the DESVD-B interaction is dependent on the ionic strength of the solution (Grove et al. 2012).

(iii) The use of the same CTPR18 protein to form a functional film by adding PEG as a plasticiser on a Teflon surface. Grove *et. al.* deposited the same CTPR18 and PEG 400 (low-molecular-weight grade PEG) on Teflon tape to allow the solvent evaporate. After solvent evaporation, solid 100μm thick multi-layered films were formed in an ordered manner (Grove, Regan, and Cortajarena 2013). In this structure, the film could still bind the DESVD peptide suggesting that some of the binding sites were "open" to binding.

(iv) A minimised designed β-roll motifs was used to form filaments through the reversible polymerisation initiated by the addition of lanthanum (Scotter et al. 2007). Scotter *et. al.* designed a 34- and 50-amino acid minimised β-roll motif based on the nonapeptide (peptide that contains nine amino acids) consensus sequence from alkaline protease. It contained three β-strands and two calcium-binding sites (34-amino acids) and an elongated version with five β-strands and five calcium-binding sites (50-amino acids). Even though both proteins are designed with calcium-binding sites, both peptides did not show any conformational change upon the addition of calcium. However, both peptides did specifically aggregate upon the addition of lanthanum (Figure 1.19B) (Scotter et al. 2007). Lanthanum is used to replace calcium because it has an additional charge and a similar ionic radius. The aggregations of these peptides caused by lanthanum were identified as ordered filaments.

(v) A β-strand was designed to form metal-mediated self-assembly β-barrel with a hollow centre. The four β-strand residues were fused with loop forming residues on the C-terminus and two pyridyl metal binding sites at the C- and N-termini (Yamagami, Sawada, and Fujita 2018). The zinc salts coordinate two β–strands into dimers, of which three dimers then self-assemble into cylindrical antiparallel β-sheets barrel with a hydrophobic pore (Figure 1.19C).

(vi) Zipper-forming protein repeats were designed to create fast self-assembling and thermally stable fibrils. Four polypeptide blocks were engineered where each block consisted of 16 repeats of a zipper-forming segment from four different amyloid morphological classes. These were separated by a flexible linker, the glycine-rich sequence from the *Nephila clavipes* MaSp1 dragline spidroin protein (Figure 1.19D) (Dai et al. 2019). All four block polypeptides were observed to self-assemble into highly ordered fibrils.

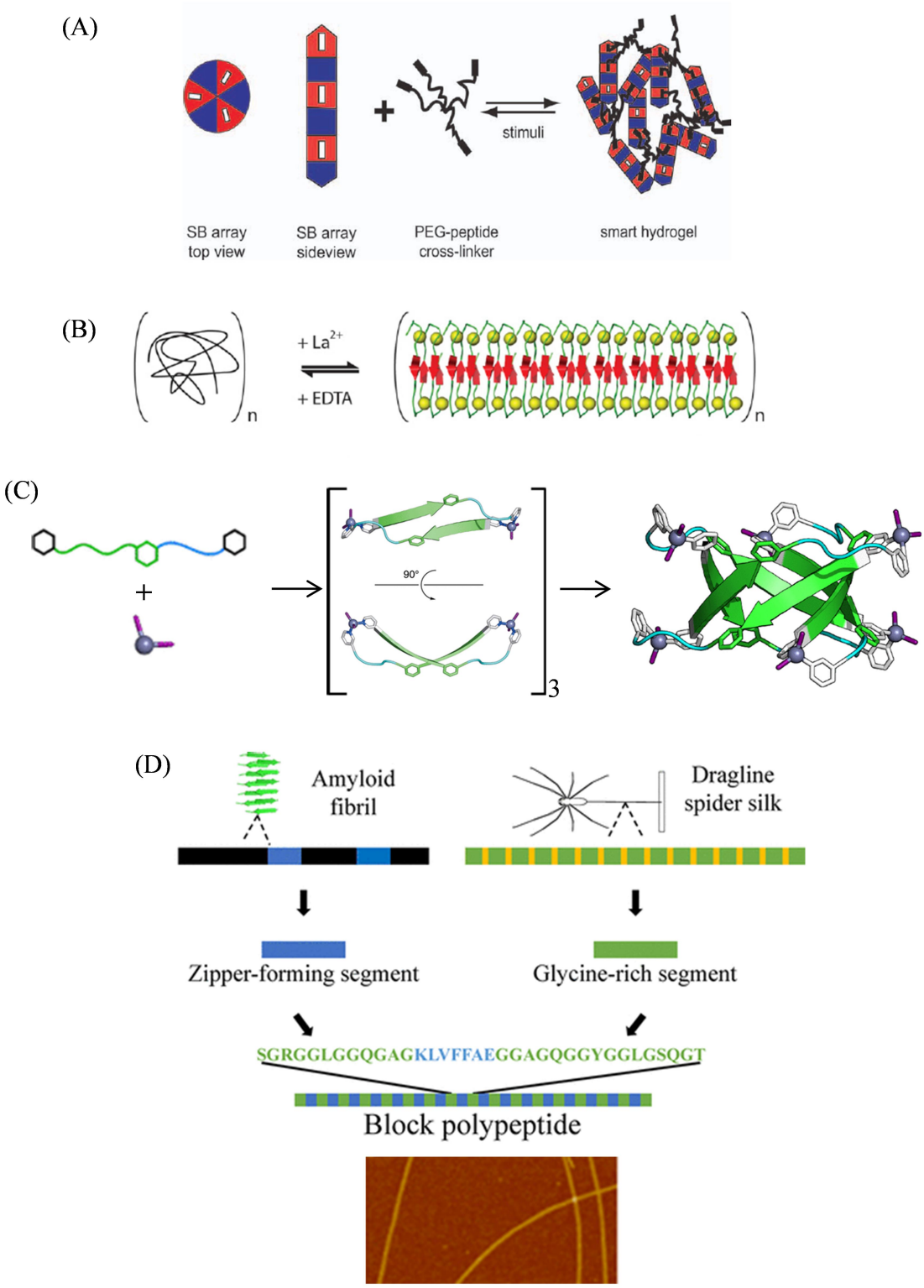


Figure 1.19 Examples of nanostructures created by repeat proteins. **(A)** Schematic diagram of the CTPR18 from top and the side view, where red blocks are CTPR18 with binding sites (B) and CTPR18 without binding sites (S). Formation of a gel with the addition of PEG-peptide cross-linker (adapted from Tijana Z. Grove *et al.*, 2012). **(B)** Schematic diagram of the reversible polymerisation of the minimised designed β-roll motifs on addition of lanthanum (adapted from Scotter *et al.*, 2007). **(C)** The formation of a synthetic β-barrel. The zinc salt (bottom molecule) induced dimer formation of the β-strand-loop fusion peptide (green represents the β-strands, blue represents the loop) and three of the dimers formed a hollow β-barrel (adapted from Yamagami, Sawada and Fujita, 2018). **(D)** Schematic diagram of the design of the polypeptide block and the atomic force microscopy image of one of the fibrils formed (bottom) (adapted from Dai *et al.*, 2019).

1.7 Native chemical ligation (NCL) driven protein assembly

Another system of protein assembly that has been employed is the native chemical ligation (NCL) mediated protein assembly. Proteins that have intein embedded in its polypeptide chain undergo NCL. Upon translation, the intein will excise itself, ligating the two flanking protein extein domains via peptide bond (Shah and Muir 2014). The formation of a peptide bond between the two exteins is known as protein splicing. During protein splicing, an N–S or N–O acyl shift forms a thioester or oxoester bond at the N-extein/intein junction Figure 1.20A. This reactive intermediate is attacked during transthioesterification by the side chain sulfhydryl or hydroxyl group of the first residue in the C-extein, which can be Cys, Ser, or Thr, to give a branched intermediate, Figure 1.20B. Then, cyclisation of the conserved Asn residue at the C-terminus of the intein releases the intein. Finally, the thioester bond between the exteins rearranges to a peptide bond by a spontaneous S–N or O–N acyl shift, Figure 1.20C.

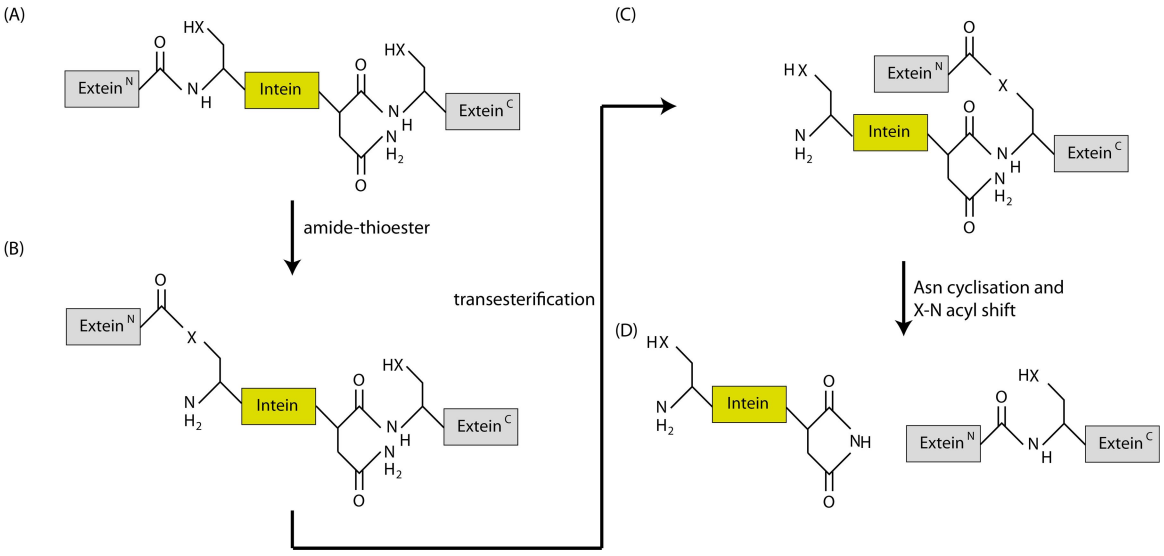


Figure 1.20 Schematic diagram of NCL driven by inteins. **(A)** Firstly, an N–S or N–O acyl shift forms a thioester or oxoester bond at the N-extein/intein junction. **(B)** This reactive intermediate is attacked in transthioesterification reaction by the side chain sulfhydryl or hydroxyl group of the first residue in the C-extein, which can be Cys, Ser, or Thr, to give a branched intermediate. (**C)** The cyclisation of the conserved Asn residue at the C-terminus of the intein releases the intein. Finally, the thioester bond between the exteins rearranges to a peptide bond by a spontaneous S–N or O–N acyl shift. X can be sulphur or oxygen (adapeted from Shah and Muir, 2014).

Here are some examples of fibre formation driven by NCL:

(i) Three repeats of CTPR (CTPR3) were used to form fibrous nanostructures by intein initiated NCL. Originally the CTPR3 was designed with an additional capping C-terminal helix to prevent oligomerisation (Main et al. 2003). This had to be removed to introduce a compatible oligomerisation interface at the C-terminus (CTPR3△S). Phillips *et. al.* fused a His-tag at the N-terminus and a Mxe Gyr A (MxGA) intein to the C-terminus of CTPR3△S (Jonathan J. Phillips, Millership, and Main 2012). The intein was modified to enable cleavage by sodium 2-mercaptoethanesulfonate (MESNa), which produced an exposed thioester at the C-terminus. Factor Xa (FXa) was then used to cleave the His-tag at the N-terminus to expose a reactive N-terminal cysteine. The NCL could then take place to produce CTPR polymerisation (Figure 1.21A) (Jonathan J. Phillips, Millership, and Main 2012).

(ii) Harvey *et. al.* demonstrated controlled fabrication of CTPR3△S by iterative NCL reaction, producing CTPR12△S with 12.5 % yield (Harvey, Itzhaki, and Main 2018). The fabrication started with the ligation of a cap, CTPR3△S with an exposed cysteine on the N-terminal (cleaved by TEV protease), and linker, CTPR3△S sandwiched by a TEV site on the N-terminus and an exposed thiol group on the C-terminal (MxGA intein cleaved by MESNa). The product can be cleaved by TEV protease to reveal a C-terminal cysteine and therefore can further ligate with another linker (Figure 1.21B). The fibres can then extend to a specific length. However, due to the ligation yield, *i.e.* 75 %, the extension of CTPR3 was limited to 4 sequential ligations joining 5 protein modules together (Harvey, Itzhaki, and Main 2018).

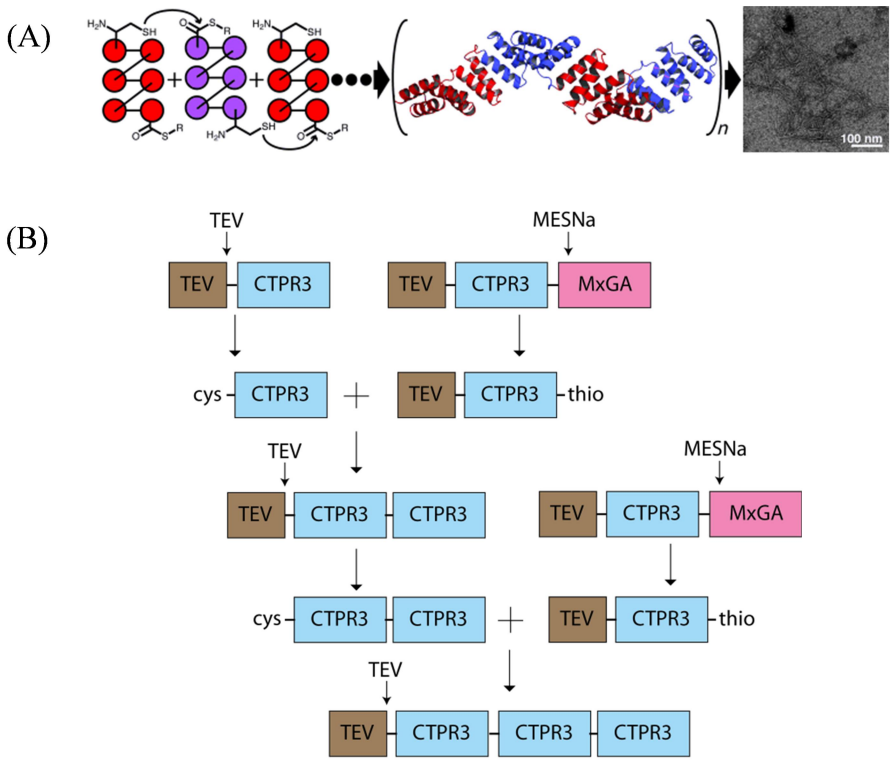


Figure 1.21 Intein-mediated protein assembly. **(A)** The fusion of CTPR3△S, His_6 and Intein. CBD is a chitin binding protein that is used for protein purification. Factor Xa and MESNa are then used to cleave the His_6 and intein, which leaves CTPR3△S to elongate with other CTPR3△S and form fibrous nanostructures, as shown in the TEM (adapted from Jonathan J. Phillips, Millership and Main, 2012). **(B)** Schematic diagram of the controlled fabrication of CTPR3△S using iterative NCL reaction via MxGA intein.

1.8 Thesis aims

As this introduction shows, many studies have shown that there are a number of ways that existing proteins can be used as templates to design novel self-assembly systems – from cages to fibres and hydrogels (Woolfson 2014; Flenniken et al. 2009; Fletcher et al. 2013; Speltz, Nathan, and Regan 2015; Grove et al. 2012; Jonathan J. Phillips, Millership, and Main 2012). Nonetheless, as can be seen, these systems are often very specific or require the user to have in depth knowledge of highly intensive computer programming. The work reported in this thesis aims to develop a generic self-assembly system that does not depend on precise protein domains and specific protein-protein interfaces. The system utilises symmetrical protein domains as building blocks and NCL mediated by intein and split-intein protein pairs as the main driving force. These design principles are applied to form different complex morphologies such as fibres and cages in a precise manner. Importantly, this work presents a novel approach for iterative protein structure assembly that is able to incorporate functionalised protein domains onto specific locations of the assembled structure.

2 Materials and Methods

2.1 Introduction

This chapter describes the materials and methods used throughout the thesis. It includes: (i) the methods used to produce recombinant DNA, (ii) expression and purification of recombinant proteins, (iii) reaction conditions used for protein ligations, (iv) purification of the ligated protein products, (v) the analytical techniques used to analyse the ligated products formed and (vi) estimation of yields.

2.2 Molecular biology

2.2.1 Vectors used

The vectors used to express the various recombinant proteins were pOPINE, pOPINF (OPPF-UK) and pET23b (Novagen) as shown in Figure 2.1. All vectors are high copy number plasmids, ampicillin resistance and are a T7 expression system where the recombinant protein was inducible with Isopropyl-β-D-1-thiogalactopyranoside (IPTG).

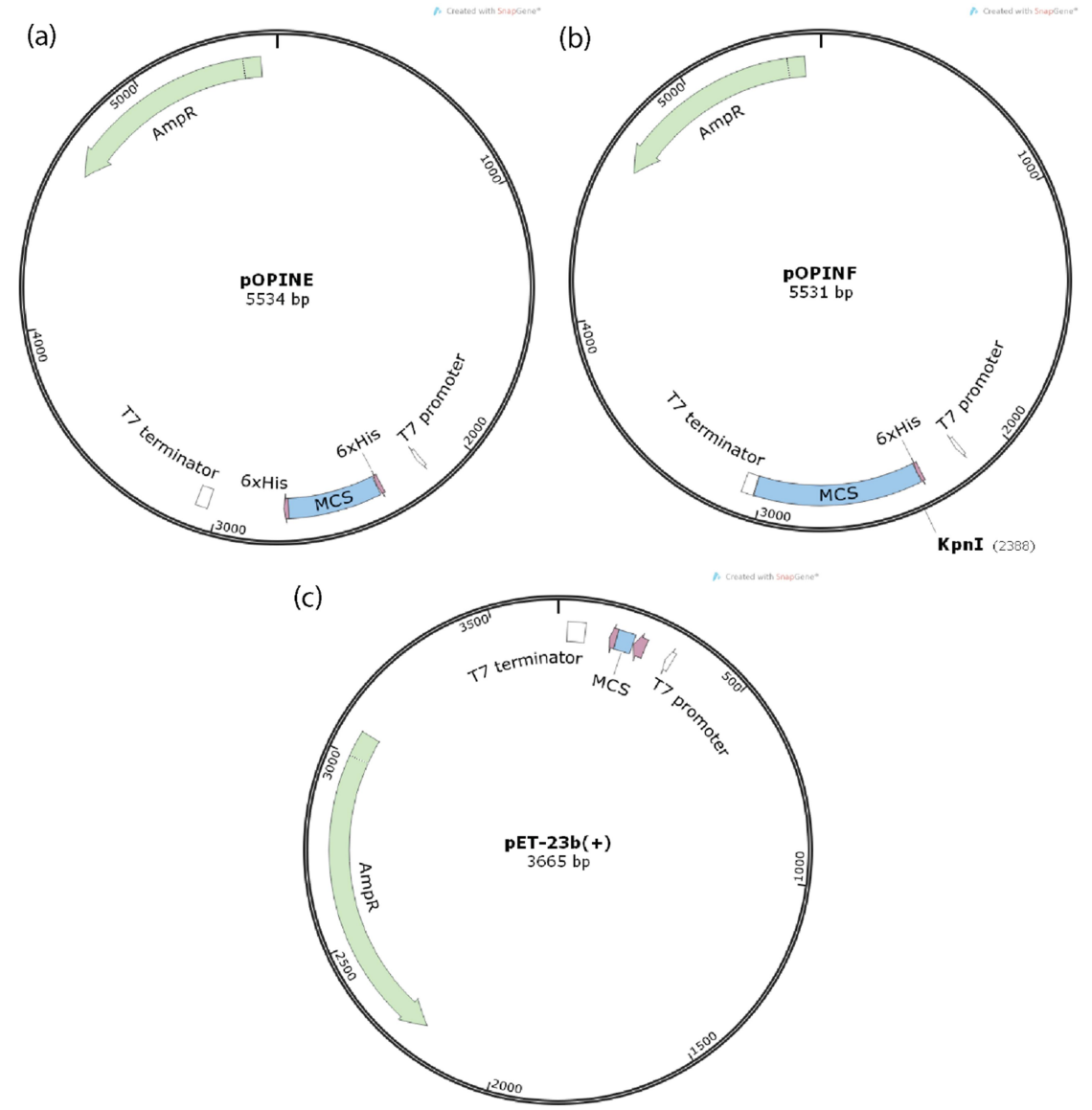


Figure 2.1 Map of the vectors used in this thesis. (a) pOPINE vector; (b) pOPINF vector; and (c) pET23b (+) vector. The maps show the AmpR is the ampicillin resistance gene (green), the multiple cloning site (MCS); the polyhistidine affinity tag (pink arrow); the T7 promoter, the T7 terminator and the KpnI site on the pOPINF vector.

2.2.2 Construction of expression vectors to produce fusion proteins genes

All recombinant fusion protein genes were constructed by cloning combinations of the protein domains listed in Table 2.1 into the vectors described in Section 2.2.1. Initially the DNA sequence encoding fusion proteins were constructed by inserting genes sequentially using the strategy outlined in Figure 2.1. First, the SpeI and BglII restriction sites were added into the expression vector along with TEV using the KpnI restriction site. Each gene was amplified, by PCR, incorporating a NheI restriction enzyme cut site 5' of the gene and SpeI and BglII restriction enzyme cut sites 3' to the gene. The expression vectors have SpeI and BglII restriction enzyme cut sites; thus to

obtain sequential cloning, the expression vector was cut with SpeI and BglII, and the amplified gene NheI and BglII restriction enzymes. The NheI and SpeI produce compatible overhangs and enable the gene to be ligated into the vector. On ligation the NheI/SpeI sites from the gene and the vector, respectively, are destroyed. This leaves the SpeI and BglII at the 3' end of the gene ready for the next gene to be ligated into. Once the 1st generation of fusion proteins were constructed, 2nd and 3rd generation fusions could be created by PCR amplifying sections of the 1st gen fusions and sub-cloning these with other appropriate restriction enzymes. Table 2.2 summarises all the fusion protein plasmids created and used. It lists the protein domains present (from N- to C-terminus). All were verified by DNA sequencing (Beckman-Coulter Genomics or Source Bioscience) and were constructed either by myself or, where indicated, by Dr. J. Wright or K. Richardson.

Table 2.1 Protein domains used in the Thesis.

Nomenclature (Abbreviation)	DNA size (bp)	Reference
Affinity Tags		
Chitin Binding Domain (CBD)	156	Watanabe et al. 1994
6-Histidine tag (H)	18	Hengen 1995
Glutathione S-transferase (GST)	218	Frangioni and Neel 1993
Inteins		
Mxe Gyr A intein (MxGA)	564	Chong et al. 1997
Gp4-1 split-intein ($Gp^{N/C}$)	264 and 111	Dassa et al. 2009
IMPDH-1 split-intein ($Imp^{N/C}$)	303 and 123	Dassa et al. 2009
Repeat Domains		
Monofoil-4P (M4P)	123	Blaber and Lee 2012
CTPR3△S (CTPR3)	306	Phillips, Millership, and Main 2012
CTPR390△S (CTPR390)	306	Cortajarena et al. 2004
Alpha Helical Linkers	30	Chen, Zaro, and Shen 2013

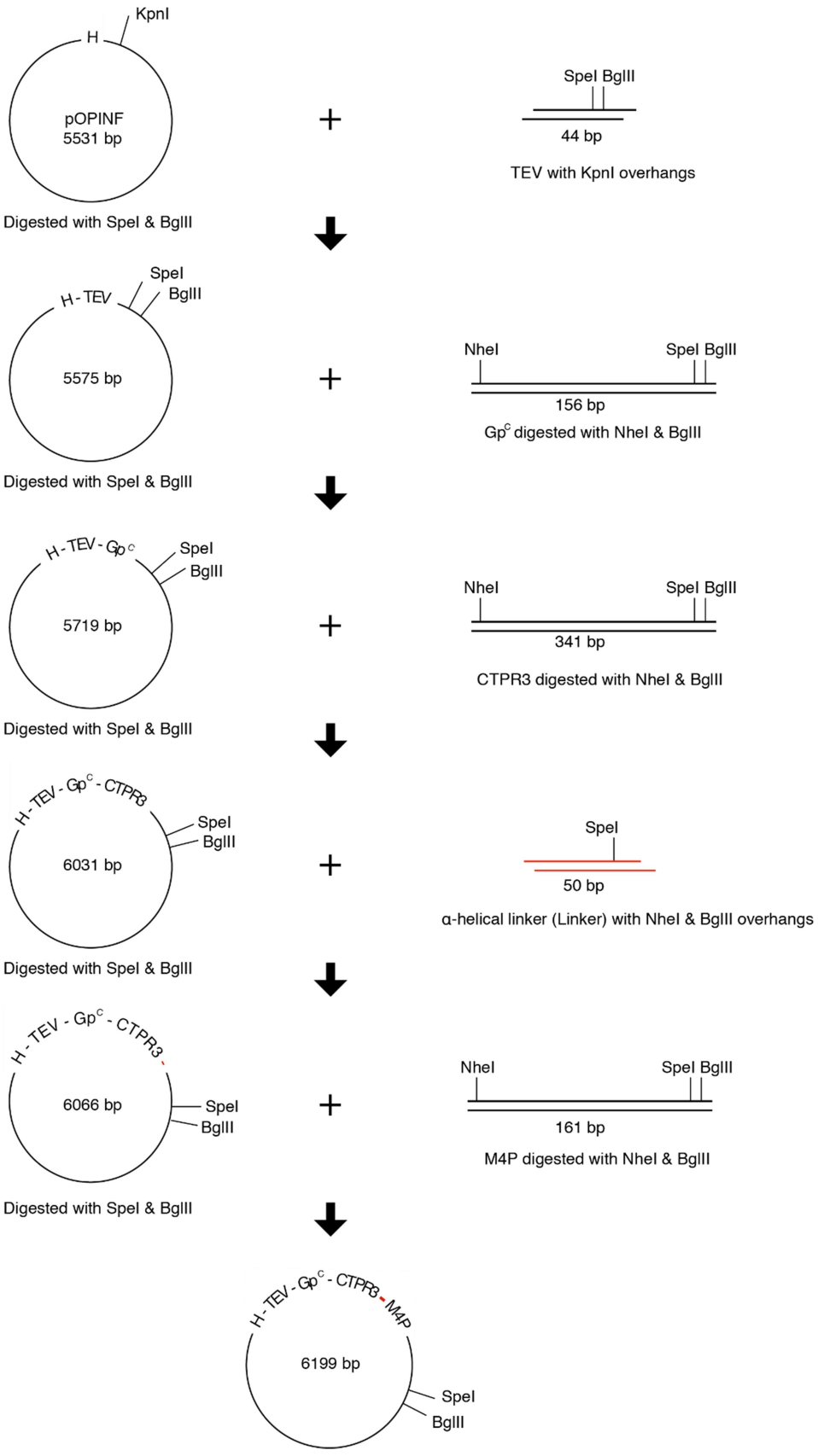


Figure 2.2 Example of the construction of fused sequences via the sequential insertion of domains. The first sequence encoding the TEV domain and SpeI/BglII restriction sites were added into the expression vector's KpnI site. Then, the following domains were added sequentially using SpeI/NheI and BglII. SpeI and NheII digestions produce compatible overhangs that can be ligated but destroy the SpeI and sites. The final product is the plasmid of the fusion protein, H-TEV-Gp^C-CTPR3-M4P.

Table 2.2 Summary of the constructs produced and chapters where they are used.

Chapter	Construct from N- to C-terminus	Constructed by
3	H-M4P-CTPR3-MxGA-CBD	JNW
3	H-TEV-CTPR3-M4P	JNW
3	H-M4P-CTPR3-Imp^N	JNW
3	H-GST-Imp^C-CTPR3-M4P	JNW
3	H-M4P-CTPR3-Gp^N	JNW
3	H-Gp^C-CTPR3-M4P	JNW
3, 5	H-CBD-Imp^C-CTPR3-M4P	WLW
3	M4P-CTPR3-Imp^N-CBD-H	WLW
3, 5	M4P-CTPR3-Gp^N-H	WLW
3	H-CTPR3-M4P	JNW
4	H-CBD-Imp^C-CTPR3-Gp^N	KR
4	H-CBD-Gp^C-CTPR3-Imp^N	JNW
4	H-CTPR3-Imp^N	JNW
4	H-CTPR3-Gp^N	JNW
4	H-Gp^C-CTPR3	JNW
4	CTPR3-Gp^N-H	WLW
4	CBD-Imp^C-CTPR3-Gp^N-H	WLW
4,5	H-Gp^C-CTPR3-Imp^N-CBD	WLW
5	H-Gp^C-CTPR390-Imp^N-CBD	WLW

Where H is the polyhistidine-tag; CBD is the chitin binding domain; GST is the Glutathione S-transferase; TEV is the Tobacco etch virus cleavage site; M4P is the Monofoil-4P; CTPR3 is the three repeats of the consensus tetratricopeptide; CTPR390 is the three repeats of the consensus tetratricopeptide with binding pocket; MxGA is the Mxe GyrA intein; Imp^N is the N-terminal domain of IMPDH-1; Imp^C is the C-terminal domain of IMPDH-1; Gp^N is the N-terminal domain of Gp41-1; Gp^C is the C-terminal domain of Gp41-1.
All constructs containing of M4P and CTPR3 / CTPR390 are connected via α-helical linkers.

2.2.3 Standard molecular biology techniques

Below outlined the commonly used molecular biology techniques employed to construct the vectors via PCR amplification and restriction digest sub-cloning described in Section 2.2.2.

2.2.3.1 Polymerase Chain Reaction (PCR)

Primer Design. In general, pairs of primers of between 20 and 40 bases long were designed with an annealing temperature of 50 – 65 °C. One primer was complementary to the coding strand and the other complementary to the non-coding strand of the gene. In addition, the appropriate restriction enzyme sequences were added to each primer so that the resultant amplified genes could be inserted into the vector in the appropriate

place. The primers were synthesised by Integrated DNA Technologies. The annealing temperature (Tm) was determined by SnapGene® software by using nearest neighbour (NN) adjusted equations as per Equation 2.1 (SantaLucia and Hicks 2004; Markham and Zuker 2008).

Equation 2.1

$$T_M = \frac{\Sigma\Delta H^{\circ}(1000)}{\Sigma(\Delta S^{\circ} + 0.368\left(\frac{N}{2}\right)(\ln[Na^{+}])) + R(In\left(\frac{C_T}{\chi}\right))} - 273.15$$

Where △H and △S values are calculated with regards to the △G thermodynamic nearest-neighbour parameters in Table 2.3, R is the gas constant (1.9872 cal/K-mol), N is the total number phosphates and [Na^+] represents the total concentration of monovalent cations present (0.05 mM), C_T is the molar strand concentration (0.25 uM) and c equals 4 for non-self-complementary duplexes and 1 for self-complementary duplexes.

Table 2.3 Nearest neighbour thermodynamic parameter for DNA Watson-Crick pairs in 1 M NaCl.

Propagation sequence	△H° (kcal/mol)	△S° (e.u.)	$\triangle G°_{37}$ (kcal/mol)
AA/TT	-7.6	-21.3	-1.00
AT/TA	-7.2	-20.4	-0.88
TA/AT	-7.2	-21.3	-0.58
CA/GT	-8.5	-22.7	-1.47
GT/CA	-8.4	-22.4	-1.44
CT/CA	-7.8	-21.0	-1.28
GA/CT	-8.2	-22.2	-1.3
CG/GC	-10.6	-27.2	-2.17
GC/CG	-9.8	-24.4	-2.24
GG/CC	-8.0	-19.9	-1.84
Initiation	+0.2	-5.7	+1.96
Terminal AT penalty	+2.2	+6.9	+0.05
Symmetry correction	0.0	-1.4	+0.43

PCR Amplification. PCR amplification was carried out in 50 µL reaction volumes containing: 10 µL Q5 5x reaction buffer; 1 µL 10 mM dNTPs, 2.5 µL 10 mM Forward Primer, 2.5 µL 10 mM Reverse Primer, 1 µL Template (<10 ng), 32.5 µL double dionised H_2O (ddH_2O), 1 µL Q5 DNA Polymerase (NEB). The reaction was initially heated to 98 °C for 2 mins to denature the DNA strands. After the initial denaturation step, 30 cycles of amplification were performed by using the following: 98 °C

denaturation for 30 secs, using the Tm of the primers to anneal for 30 secs and 72 °C extension for 30 secs per 1 kbp. The PCR ended with a final extension, 72 °C for 5 mins and the PCR product was stored at 4 °C.

2.2.3.2 DNA digestion

Plasmid and PCR product digestion was carried out with the following components: 1 µg DNA, 2 µL 10x FastDigest Green buffer, 1 µL of each FastDigest restriction enzyme (Thermo Fisher Scientific) and ddH_2O to a total volume of 20 µL. The reaction was mixed and incubated at 37 °C for 30 mins to 1 hr.

2.2.3.3 DNA gel electrophoresis

PCR and digested products were analysed and purified (when required) using DNA electrophoresis. A 1% agarose gel was made by dissolving 0.5 g of agarose powder (Sigma) in 50 mL TAE buffer. Ethidium bromide was added to a final concentration of 0.04 % and the gel was allowed to set at 25 °C. Samples were loaded into wells after being mixed with 6x loading dye or premixed with FastDigest Green buffer (Thermo-Fisher). A marker, Quickload 2 -log DNA ladder (NEB) was used and the gel was run at 80 V for 30 mins. Gels were visualised using a UV transilluminator.

2.2.3.4 DNA clean-up and gel extraction

PCR and digested products were extracted from the gel using Monarch PCR and DNA Clean-Up Kit (NEB) following the manufacturer's instructions. When there were multiple PCR products or extracting digested DNA product was required, the wanted products were extracted from the DNA Gel after separating DNA fragments via gel-electrophoresis.

2.2.3.5 DNA ligation

DNA ligation reactions contained the following components: 2 µL 10 x T4 DNA Ligase buffer, 1 µL T4 DNA ligase, digested insert DNA, digested vector DNA (100 ng) and ddH_2O to a total volume of 20 µL. Ligations were performed with insert to vector ratios

of 1:1, 3:1 and 7:1. 100 ng of cut vector was always used with the mass of insert calculated as follows:

Equation 2.2

$$\text{Mass Insert (ng)} = \left(\frac{InsertRatio}{VectorRatio}\right) \times Mass\ Vector\ (ng) \times \left(\frac{InsertLength(bp)}{VectorLength(bp)}\right)$$

In the case of blunt-ended ligations the reaction was supplemented with 2 µl of 50 % PEG 4000 to aid efficiency. The reactions were carried out at 22 °C for 60 mins, followed by heat inactivation at 75 °C for 5 mins.

2.2.3.6 Bacterial transformation of ligation reactions

For all molecular biology, the *E. coli* strain XL-2 Blue (genotype: *endA*1*, supE*44*, thi*-1, *recA*1*, gyrA*96*, relA*1*, lac*, Agilent) was used. 50 µL of cells were thawed on ice before mixing with 1 µL of DNA and transferred to a 2 mm electroporation cuvette (Cell Projects, UK). The prepared cuvette was transferred to a Bio-Rad Micropulsar and pulsed once at 2.5 kV. If the time constant recorded > 3 ms the electroporation was considered successful. 500 µL of SOC media (2 % Tryptone, 0.5 % Yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 20 mM glucose, 10 mM $MgSO_4$) was added to the cuvette and mixed with the cell suspension before being transferred to a 7 mL bijou container and incubated at 37 °C, 200 rpm, for 1 hr. All the recovered transformation was plated onto a LB agar plate supplemented with 100 µg/mL of ampicillin. Plates were incubated at 37 °C overnight before being stored at 4 °C.

2.2.3.7 Colony PCR screening

To obtain XL2 colonies with the desired ligation, PCR screening was carried out. A pre-made buffer and polymerase mix [OneTaq Quick Load Master Mix (NEB)] with primers designed to amplify over the insert region of the vectors were used. Aliquots of 5 µL of 2x OneTaq Quick Load Master Mix, 0.25 µL of 10 µM Forward primer, 0.25 µL of 10 µM Reverse primer and 4.5 µL of ddH_2O were placed into 96 well plates. To each well a single XL2 colony was mixed and then plated out. The PCR reactions were then reacted using the following cycle: initial denaturation 94 °C, 30 secs; 30 cycles of 94 °C, 15 secs; 48 °C, 15 secs; 68 °C, 1 min per kb; final extension 68 °C, 5 mins.

Reactions were then loaded onto a DNA agarose gel (prepared as described above) and positive colonies selected for DNA plasmid extraction.

2.2.3.8 DNA plasmid extraction

Colonies selected for plasmid DNA extraction were prepared by inoculating single colonies into separate 10 mL aliquots of LB supplemented with 100 µg/mL ampicillin. Cultures were incubated overnight at 37 °C, 200 rpm. 4 mL of each culture was harvested in 2 mL Eppendorf tubes (repeated to increase the pellet size). Plasmid DNA was extracted from the harvested cells using the GeneJet plasmid miniprep kit (ThermoFisher Scientifc). DNA quality was assayed by gel electrophoresis and spectrophotometry.

2.2.3.9 Sanger sequencing

To confirm the sequence of the constructs, DNA plasmids extracted from positive colonies were sent for Sanger Sequencing. Sequencing was carried out by BeckmanCoulter Genomics or Source Bioscience, using T7 Promoter primer or M13-Forward primer. Sequencing quality was assessed and aligned to expected sequence results using SnapGene Software (GSL Biotech).

2.2.3.10 Preparation of electro-competent cells

Electro-competent *E. coli* cell lines C41 (DE3) (Lucigen) and XL2-Blue (Lucigen) were used in this thesis. XL2-Blue cells were used for molecular biology as they improve insert stability due to the deficiency of the recombination (*recA*) gene. They prevent the cleavage of cloned DNA by the EcoK endonuclease system due to the hsdR mutation. XL2-Blue cells also improve the quality of miniprep DNA, as they are endonuclease (*endA*) deficient. *E. coli* strain C41 (DE3) electro-competent cells were used for recombinant protein expression in this thesis. *E. coli* C41 cells are derived from BL21 (DE3). It has a mutation to prevent cell death when overexpressing a variety of toxic proteins. C41 contain the T7 RNA Polymerase gene needed to produce the T7 promoter for vector expression. The T7 RNA Polymerase gene is under the control of the LAC promoter; therefore, expression can be controllably induced by addition of IPTG.

Both electro-competent cells were produced in our laboratory. The following protocol was used for both cells lines. As XL2-Blue cells are resistant to tetracycline, at a working concentration of 12.5 µg / mL, tetracycline (Sigma-Aldrich) was added to LB agar plates and growth media. LB agar plates and media were not supplemented with an antibiotic for C41 (DE3) cells. The desired *E. coli* strain was streaked from a glycerol stock onto LB agar plates and incubated overnight at 37 °C. 10 mL of LB media was inoculated from a single colony taken from the streaked plates and incubated overnight at 37 °C, at 200 rpm. 500 mL 2xYT cultures were inoculated from the 10 ml starter culture and grown until an optical density at 600 nm wavelength (OD_{600}) of 1.0 was reached. Cultures were immediately chilled on ice for 30 mins. Cells were harvested by centrifugation at 5000 rpm in an Eppendorf table top centrifuge for 10 mins at 4 °C and re-suspended in 500 mL of sterile ice cold 10 % glycerol. Cells were pelleted again by centrifugation at 5000 rpm for 10 mins at 4 °C and re-suspended in 250 mL of sterile ice cold 10 % glycerol. This process was repeated for 100 mL and finally 1 mL of ice cold 10 % glycerol. Cells were then resuspended in a minimal amount of 10 % glycerol and 50 µL aliquots flash frozen in liquid nitrogen. These were stored at -80 °C until use.

2.3 Recombinant protein expression and purification

2.3.1 Transformation of *E. coli* C41 (DE3) electro-competent cells with recombinant plasmids

E. coli C41 (DE3) electro-competent cells were transformed with the recombinant plasmids using the same procedure described in Section 2.2.3.6. 50 µL of the transformants were plated on a LB agar plate supplemented with 100 µg/mL of ampicillin. Plates were incubated at 37 °C overnight before being stored at 4 °C for up to 4 weeks.

2.3.2 Protein expression

A single colony of transformed *E. coli* C41 electro-competent cells were picked into 100 mL LB starter culture supplemented with 100 µg/mL ampicillin. This was incubated for 16 hr at 37 °C, 200 rpm. 1 L of 2xYT media supplemented with 100 µg/mL of ampicillin was inoculated with 10 mL of the starter culture and grown at

37 °C until an OD_{600} of ~1.0. Expression was induced by the addition of IPTG to a final concentration of 1 mM over a range of temperatures and times that were construct dependent. Table 2.4 summarises the protein expression conditions for each construct used.

2.3.3 Protein purification

To purify the proteins, both native and denaturing methods were employed. These used both affinity chromatography and size exclusion chromatography (SEC) as required. Table 2.4 summarises how each construct was purified with Sections 2.3.3.1 to 2.3.3.3 describing each type of purification in detail.

Table 2.4 Summary of the expression conditions, method of purification and yield for each protein described here.

Chapter	Construct name	Expression condition		Method of purification	Yield (mg L^{-1})
		Temp. (°C)	Duration		
3	H-M4P-CTPR3-MxGA-CBD	37	3 hr	Chitin	8.4
3	H-TEV-CTPR3-M4P	37	Overnight	Denat.	50
3	H-M4P-CTPR3-Imp^N	37	3 hr	Denat.	19.5
3	H-GST-Imp^C-CTPR3-M4P	37	4 hr	Denat.	39.5
3	H-M4P-CTPR3-Gp^N	37	Overnight	Denat.	12
3	H-Gp^C-CTPR3-M4P	37	Overnight	Denat.	25
3, 5	H-CBD-Imp^C-CTPR3-M4P	37	4 hr	Denat.	15.0
3	M4P-CTPR3-Imp^N-CBD-H	16	Overnight	Native	20.5
3, 5	M4P-CTPR3-Gp^N-H	16	Overnight	Denat.	10
3	H-CTPR3-M4P	16	Overnight	Native	13.6
4	H-CBD-Imp^C-CTPR3-Gp^N	37	Overnight	Denat.	20.75
4	H-CBD-Gp^C-CTPR3-Imp^N	37	Overnight	Native /[a]Denat.	41.7
4	H-CTPR3-Imp^N	16	Overnight	Native	15
4	H-CTPR3-Gp^N	16	Overnight	Native	15
4	H-Gp^C-CTPR3	37	Overnight	Denat.	20
4	CTPR3-Gp^N-H	16	Overnight	Native	40
4	CBD-Imp^C-CTPR3-Gp^N-H	37	Overnight	Denat.	55
4, 5	H- Gp^C-CTPR3-Imp^N-CBD	16	Overnight	Native /[a]Denat.	50.0
5	H-Gp^C-CTPR390-Imp^N-CBD	16	Overnight	Denat.	53.5

[a]In these cases the denaturing purification yielded higher purity than the native purification.

2.3.3.1 Native expression and nickel affinity purification

Cells were induced and grown 16 °C overnight. Cells were harvested by centrifugation for 10 mins at 10,000 *x g*, the cell pellet was resuspended in 50 mM Tris pH 8, 300 mM NaCl (equilibration/wash buffer) and snap frozen in liquid nitrogen before being stored at -80 °C. The resuspended cell pellets were thawed and lysed by sonication on ice for 10 mins (per 100 mL) with 30 secs pulse using a Vibra-cell sonicator (Sonics). Insoluble matter was removed by centrifugation for 30 mins at 39,000 *x g* at 4 ^{o}C.

The cleared protein lysate was loaded onto a pre-equilibrated Nickel-iminodiacetic acid (Ni-IDA) resin. To charge the resin, 10 to 20 mL of iminodiacetic acid (IDA) resin was incubated with 2 column volumes (CVs) of 100 mM $NiSO_4$ in a 50 mL gravity flow column (Pierce). It was then washed with 5 CVs of ddH_2O and equilibrated with 2 CVs of equilibration buffer prior to loading the clear protein lysate. Once the lysate had been loaded onto the Ni-IDA resin, it was washed with 10 CVs of equilibration/wash buffer. The recombinant protein was eluted from the column with equilibration/wash buffer with addition of 250 mM imidazole. Dithiothreitol (DTT) was added to the elution with a final concentration of 5 mM. To verify the protein purified and check the purity, elution fractions were analysed with sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). If further purification was required, size exclusion chromatography (SEC) was performed using a HiLoad 16/60 Superdex 75 or 200 prep grade on a fast protein liquid chromatography (FPLC) system (GE Healthcare).

2.3.3.2 Denatured expression and nickel affinity purification

After induction, the temperature was maintained at 37 ^{o}C for between 4 hrs and overnight (refer to Table 2.4). Cells were harvested by centrifugation for 10 mins at 10,000 *x g*. The cell pellet was resuspended in 50 mM Tris pH 8, 6 M GuHCl, 300 mM NaCl (equilibration and wash buffer 2) and snap frozen in liquid nitrogen before being stored at -80^{o}C. The resuspended cell pellets were thawed and lysed by sonication on ice for 10 mins (per 100 mL) with 30 sec pulses. Insoluble matter was removed by centrifugation for 30 mins at 39,000 *x g* at 4^{o}C.

The cleared protein lysate was then loaded onto a pre-equilibrated Ni-IDA resin and washed with equilibrated/wash buffer containing 6 M GuHCl. The resin was charged as described in previous section. Proteins were eluted unfolded in 50 mM Tris pH 8, 6 M GuHCl, 300 mM NaCl and 250 mM imidazole or refolded via stepping down the

denaturant concentration (3 M GuHCl, 1.5 M GuHCl, 0.75 M GuHCl and 1 M urea) in the wash buffer 2 whilst bound to the Ni-IDA resin. When eluted denatured, proteins were refolded via dialysis with SnakeSkin 3.5K MWCO into 50 mM Tris pH 8, 1 M urea, 300 mM NaCl and 5 mM DTT. The refolded protein on the column was eluted with 50 mM Tris pH 8, 1 M urea, 300 mM NaCl and 250 mM imidazole. DTT was added to the elution with to a final concentration of 5 mM. If further purification was required, SEC was performed using a HiLoad 16/60 Superdex 200 prep grade attached to a FPLC system. To verify the protein purified and check the purity, elution fractions were analysed by SDS-PAGE.

2.3.3.3 Native expression and chitin affinity purification

Recombinant proteins that contain chitin binding domains attached to full intein domains were, when required, purified using chitin resin. An example of this is the fusion protein H-M4P-CTPR3-MxGA-CBD. For expression, cells were incubated for 4-6 hr at 25 °C after induction and then harvested by centrifugation for 10 mins at 10,000 *x g*. The cell pellet was resuspended in 50 mM Tris pH 8, 150 mM NaCl (equilibration/wash buffer 3) and snap frozen in liquid nitrogen before being stored at -80 ^{o}C. The resuspended cell pellets were thawed and lysed by sonication on ice for 10 mins (per 100 mL) with 30 secs pulse. Insoluble matter was removed by centrifugation for 30 mins at 39,000 *x g* at 4 ^{o}C. The clarified lysate was loaded onto 20 mL of Chitin Beads (NEB) in a 50 mL gravity flow column after the resin had been prepared with 2 CVs of equilibration/wash buffer 3. After washing with 3 CVs of equilibration/wash buffer 3, the fusion protein was eluted via intein-mediated cleavage with 1 CV of cleavage buffer (50 mM Tris, pH 8, 100 mM NaCl, 10 % sodium 2-mercaptoethanesulfonate (MESNA)). The cleavage buffer was added into the column and incubated at 25 °C for 16 hrs. After cleavage, the protein containing a C-terminal thioester (H-M4P-CTPR3-thio), was washed from the column. The CBD and intein was left bound to the resin. To verify the protein purified and check the purity, flow-through fractions were analysed with SDS-PAGE.

2.3.3.4 Further purification by Size Exclusion Chromatography (SEC)

To further purify proteins, SEC was carried out using either a S75 or S200 HiLoad FPLC Superdex 16/60 filtration column (dependant on the size of the recombinant protein and the nature of contaminant). The Superdex column was attached to an AKTA Pure chromatography system (GE Healthcare), utilising a flow-rate controller that kept the pre-column pressure to < 0.5 mPa. The column was equilibrated with appropriate filter sterilised (0.2 μM) buffer prior to loading. Up to 5 mL of ~ 100 μM protein was loaded onto the column via the sample pump (0.5 ml per min flow rate). The column was then run using appropriate buffer and 4 mL elution fractions were collected. The eluted fractions were analysed with by SDS-PAGE and the elution fractions with similar purity were pooled.

2.3.4 Protein concentration

Pure recombinant proteins were concentrated to 100-200 μM with Amicon Ultra-15 (Merk-Millipore) 3K MWCO centrifugal filter units. Protein concentrations and yield were determined from the absorbance at 280 nm wavelength using the extinction co-efficient (ε) calculated from the amino acid sequence as per Equation 2.3:

Equation 2.3

$$\varepsilon(M^{-1}cm^{-1}) = (No.of\ Trp \times 5500) + (No.of\ Tyr\ \times 1490) + (No.of\ Cys \times 125)$$

The intensity of absorbance was scanned from 200 nm to 400 nm wavelength using a WBA Biowave II UV Spectrophotometer in a 1 cm path length Quartz cuvette. The concentration of the protein was calculated with Equation 2.4.

Equation 2.4

$$C = \frac{(Abs_{280} \times 1000000)}{(\varepsilon \times l)}$$

where C is the concentration in μM, Abs_{280} is the absorbance at 280 nm, ε is the extinction coefficient in $M^{-1}cm^{-1}$ and l is the path length of the cuvette in cm.

2.3.5 Protein storage

All recombinant proteins purified in this thesis were stored in 1 mL aliquots at -80 °C after being flash frozen using liquid nitrogen. In general, they were stored in 50 mM Tris pH8, 300 mM NaCl, 5 mM DTT, 0-1 M urea.

2.4 Native chemical ligation

2.4.1 Native chemical ligation via Mxe Gyr A intein.

For NCL to take place via Mxe Gyr A intein, a C-terminal thioester protein and a N-terminal cysteinyl protein were required. These were achieved by the following:

2.4.1.1 Activation of C-terminal thioesther.

C-terminal thioester formation of H-M4P-CTPR3-MxGA-CBD was achieved by induced cleavage of the C-terminal Mxe GyrA intein using the reducing agent MESNa as described in Section 2.3.3.3

2.4.1.2 Activation of N-terminal cysteine

The H-TEV-CTPR3-M4P N-terminal activation was achieved via Tobacco Etch Virus (TEV) protease cleavage of the N-terminal His-tag. The TEV protease cleavage site used is the mutated sequence: ENLYFQ↓C, rather than the commonly used site ENLYFQ↓G. Thus, when TEV cleaves, it reveals an N-terminal cysteine. Purified H-TEV-CTPR3-M4P was concentrated to 100 μM and cleaved at 25 °C in 50 mM Tris pH 8, 150 mM NaCl, 5 % (v/v) glycerol, 5 mM tris(2-carboxyethyl)phosphine (TCEP) for 16 hrs. Cleaved N-cysteinyl protein was then purified with Ni-IDA from uncleaved protein.

2.4.1.3 Reaction of the activated components.

The ligation of the two activated homo-trimers, H-M4P-CTPR3-thioesther and cysteinyl-CTPR3-M4P, was performed in 50 mM Tris-HCl pH 8.5, 1 M NaCl, 200 mM MESNa, 10 mM TCEP, for 24 hrs at 30 °C with a 1:1 molar ratio of protein and final protein concentrations of 50 μM. The reaction was also attempted with the presence of denaturant *i.e.* 4 M GuHCl / urea. The results of ligations were analysed using SDS-PAGE.

2.4.2 Native chemical ligation reactions via split-inteins

All ligations, unless otherwise stated, were carried out in mild conditions using a reaction buffer (50 mM Tris pH 8, 300 mM NaCl, 5 mM DTT) with different concentrations of urea and incubated at room temperature (25 °C). The reactions were completed either: (1) on affinity column (chitin beads) or (2) in solution. The reactions for both methods were left to proceed under mild agitation. The ligated product can be purified either with Ni-IDA or chitin beads (IBA Lifesciences) depending on the affinity tags present on the reactants and products. Further purification was done with SEC.

2.5 Protein analysis

2.5.1 Denaturing SDS-PAGE

Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE) was used to determine protein purity and to monitor ligation reactions. Protein samples were mixed with 2x loading buffer where the final buffer conditions were: 62.5 mM Tris-HCl, 2 % SDS (w/v), 10 % glycerol (w/v), 100 mM DTT, 0.01 % bromophenol blue (w/v). 2-20 μL of sample were loaded onto either 14 or 18 % PAGE gels depending on the expected protein sizes. The gel was run at 180 V for 50 - 65 mins, in running buffer: 25 mM Tris-HCl (pH 8.3), 192 mM glycine, 1 % (w/v) SDS. Bands were visualised by staining using Coomassie Brilliant Blue R-250 and subsequently destained in 10 % (v/v) acetic acid 10 % (v/v) methanol. The resulting gel was digitised using a ChemiDoc Touch (Bio-Rad). Images were then annotated using associated Image Lab software (Bio-Rad).

2.5.2 Reaction yield from SDS-PAGE gels

Protein bands were visualised using coomassie dye and yields calculated using an Odyssey Li-COR in 800 nm imaging channel. Integrated intensity values (I) corresponding to each protein band were thereby obtained.

The equation below was used to obtain the percentage of ligated product formed:

Equation 2.5

$$\% \text{ Yield} = \left[\frac{\left(\frac{I_P}{MWt_P}\right)}{\left[\left(\frac{I_P}{MWt_P}\right)+\left(\frac{I_R}{MWt_R}\right)\right]}\right] \text{ x } 100$$

where I_P is the integrated intensity of the ligated product, MWt_P is the molecular weight of the ligated product, I_R is the integrated intensity of the most consumed reactant and MWt_R is the molecular weight of the most consumed reactant. Equation 2.5 assumes that the binding of coomassie stain (and, therefore, the intensity) is linearly related to the molecular weight of each NCL protein.

2.5.3 Analytical Size Exclusion Chromatography (SEC).

Analytical SEC was carried out using the Superdex^{TM} 200 10/30 for the half cage caps, cages formed from their ligation and the linkers. For extended cage caps and extended cages, a Superose^{TM} 6 10/30 column was used. The AKTA Pure or Purifier systems (GE Healthcare) were used to operate the columns. The columns were equilibrated into running buffer (50 mM Tris pH8, 300 mM NaCl, 5 mM DTT, 0 - 2 M urea (as present in protein buffer)). 100 µL of protein sample (50 - 100 µM) was loaded by loop onto the column and run at a pressure-controlled flow rate (< 1.4 mPa) with a maximum flow rate of 1.5 mL min^{-1}. 1.2 CVs of running buffer were run. In certain cases, fractions were collected in 0.4 ml aliquots and analysed by SDS-PAGE to identify peaks and purity. UV absorption peaks were processed using the Unicorn software (v5.0, GE Healthcare) and analysed for polydispersity, before being exported and plotted using Excel (Microsoft Corporation). UV 280 nm signals were normalised to a maximum of 50 mAU for clarity. Size was estimated using the Amersham low molecular weight gel filtration calibration kit containing the following standards: albumin 67 kDa, ovalbumin 43 kDa, chymotrypsinogen A 25 kDa. To quantify the relative position of each peak the V_e/V_o was calculated and plotted against the $\log_{10}$ of the molecular weight (kDa). A

linear trend line was drawn, and the equation used to calculate the molecular weights of unknown samples. Where V_e is the elution volume (mL) at the maxima of the UV absorption peak appears; and V_o is the void volume (mL) of the mobile phase.

2.5.4 Western Blot

SDS-PAGE gels was performed as Section 2.5.1 and transferred to the membrane using the Bio-Rad Trans-Blot® Turbo™ Transfer System using its preset mixed molecular weight program (1.5 A, 25 V for 7 mins). After blocking with 5 % milk in 1x PBS, 0.1 % Tween (PBST) for 1 hr, the membrane was washed twice with PBST and incubated in PBST with the primary antibodt at 1:1000 (monoclonal anti-His/CBD Tag antibody produced in mouse (Sigma)) at room temperature for 1 hr. Membranes were washed twice in PBST for 5 mins and incubated with 1:20000 PBST of IRDye® 680LT Goat anti-Mouse IgG in PBST for 1 hr in room temperature. Blots were then washed with PBST and PBS before being imaged with a Li-COR Odyssey Infrared Imaging System.

2.5.5 Mass spectrometry

Matrix assisted laser desorption/ionisation - time of flight mass spectrometry (MALDI-TOF/MS) was carried out to determine accurate masses of recombinant proteins and post-splice reaction. To remove/reduce buffer components, samples were prepared either by (i) dilution using 50% acetonitrile, 1 % or 0.1 % trifluoroacetic acid in water (Sigma) or (ii) using EMD Millipore Zip-Tip® pipette tips according to the manufacturer's instructions. 1 µl of sample from either dilution or Zip-Tip® was mixed 1:1 with saturated sinapinic acid matrix (dissolved in 50% acetonitrile), spotted onto a Bruker MALDI-TOF/MS steel plate and allowed to air dry. The target plate was loaded into the Bruker 2000 MALDI-TOF mass spectrometer. Using positive ion mode, the gain and laser power was adjusted until the optimal signal/noise ratio was achieved. The TOF was operated in the reflectron or linear mode and each spectrum was an average of 500 laser shots. The spectra were calibrated using Protein standards 1 & 2 (Bruker). The data was extracted and visualised using R, Bruker Flex Analysis or Microsoft Excel software.

2.5.6 Circular dichroism

Circular dichroism was used to compare the number of α-helices present in the precursor proteins and the final product from the NCL reaction. Protein concentrations of 2 – 10 μM in 10 mM Tris pH 8, 50 mM NaCl and 2 mM TCEP were analysed in a 0.5 mm path length cuvette using a Chirascan™ CD Spectrometer (Applied Photophysics Ltd, UK). For each sample, a spectrum from 190 – 280 nm was recorded with points taken at 1.0 nm intervals and 0.5 secs per point scanning time. The averaged spectrum of 3 repeats was taken for each sample. Data was converted to molar ellipticity using the equation below,

Equation 2.6

$$\theta_{molar} = \frac{100 \times \theta_{obs}}{M \times l}$$

where θ_{molar} is the molar ellipticity in deg cm^2 dmol^{-1}, θ_{obs} is the observed CD signal in millidegrees, l is the path length in cm and M is the molar concentration of protein.

2.5.7 Size exclusion chromatography small angle x-ray scattering (SEC-SAXS)

SAXS cage samples were prepared by ligating 100 mL of 10 μM IMPDH tagged half cage caps and purified using the same affinity and size exclusion chromatography steps outlined in Section 2.3.3. Purified cages were concentrated to 10 mg/ml and dialysed into 10 mM Tris pH8, 50 mM NaCl, 5mM DTT buffer. SAXS experiments were recorded on beamline B21 at the Diamond Light Source (DLS), UK, coupled to a Shodex KW403-4F size exclusion column. Data were measured at 20 °C with a wavelength of 0.99 Å and a 3 s exposure time per frame on a Pilatus 2 M two-dimensional detector at 4.014 m distance from the sample, corresponding to a momentum transfer range of $0.004 < q < 0.4$ Å^{-1} ($q = 4\pi \sin\theta \lambda^{-1}$, 2θ is the scattering angle).

2.5.7.1 SEC-SAXS analysis

Elution peak, buffer selection, and subsequent buffer subtraction, intensity normalisation, and data merging were performed in ScÅtter (BIOISIS). Further analysis was carried out with a q range of $0.018 < q < 0.35$ Å^{-1}. The radius of gyration (Rg) and

scattering at zero angle (I(0)) were calculated from the analysis of the Guinier region by AUTORG (Petoukhov et al. 2007). AUTORG is a command-line program to estimate the radius of gyration (Rg) using the Guinier approximation:

Equation 2.7

$$I(s) = I(0)\exp(\frac{S^2Rg^2}{-3})$$

Where I(s) is the scattering intensities, I(0) is the scattering intensities at zero angle, S is the scattering vector and Rg is the radius of gyration.

The value of Rg is estimated from the best possible linear fit of ln[I(s)] versus S^2 (Guinier plot), which is valid for sufficiently small scattering vectors (in the range up to sRg ≤ 1.0–1.3). The radius of gyration provides an estimate of the overall size of a particle (the root-mean-square distance to centre-of-mass in a particle).

The distance distribution function (P(r)) was subsequently obtained using GNOM (D. I. Svergun 1992), yielding the maximum particle dimension (D_{max}). It reads in one-dimensional scattering curves and evaluates the particle distance distribution function p(r). To demonstrate the absence of concentration-dependent aggregation and interparticle interference Rg over the elution peaks was inspected and analysis performed on frames where Rg was the most stable. The Porod exponent and molecular weight were calculated within the ScÅtter (BIOISIS) and ATSAS package (D. Franke et al. 2017), respectively. The Porod exponent is to determine the flexibility of the protein.

2.5.7.2 *Ab initio* shape determination from SAXS

The *ab initio* modelling was done using either GASBOR or DAMMIF (D. I. Svergun, Petoukhov, and Koch 2001; Daniel Franke and Svergun 2009) averaging 100 simulations. GASBOR reconstructs protein structure by a chain-like ensemble of dummy residues corresponding to average residue densities which placed anywhere in continuous space with a preferred number of close distance neighbours for each atom. The centres of these residues aim to approximate positions of the C-α atoms in the protein structure. In addition, the number of residues equal that in the protein. DAMMIF is a bead modelling program. In bead modeling, a particle is represented as a collection of a large number of densely packed beads inside a search volume. Each bead

belongs either to the particle or to the solvent. Starting from an arbitrary initial model DAMMIF utilises simulated annealing to construct a compact interconnected model yielding a scattering pattern that fits the experimental data. The solutions produced by GASBOR/DAMMIF were shortlisted based on biophysical data and averaged using DAMAVER (Volkov and Svergun 2003). The averaged model was used as a template to generate a DAMMIN model (D. I. Svergun 1999).

2.5.7.3 Comparison of experimental SAXS profile to generated atomic cage models

The SAXS profile was compared to 30 manually generated differing atomic models of possible designed cage conformations using the program Crysol (D. Svergun, Barberato, and Koch 1995). The models were constructed by first manually positioning the crystal structure available in the Protein Data Bank (PDB) using PyMol. COOT was used to add connecting sequences between the domains using its Rigid Body Fit Zone functionality (Emsley et al. 2010) and three-fold symmetry enforced using PyMol. The chains were renumbered (PDBSet, CCP4 suit (Collaborative Computational Project, Number 4 1994)) and rigid body refined with REFMAC5 (Vagin et al. 2004) to obtain the final lowest energy confirmation. These final lowest energy structures were converted to a SAXS profile by Crysol and compared to the experimentally determined profile. Crysol evaluates the solution scattering from macromolecules with known atomic structure and fits it to experimental scattering curves from SAXS.

3 Design of Self-Assembled Cage Structure

3.1 Introduction

This chapter explores how recombinant protein fusions can be designed to self-assemble into protein cages/encapsulations. Geometric symmetry design was used to specify the desired shapes, with assembly driven by genetically encoded native chemical ligation. To validate our method and investigate the limits of the system, trimeric half-cage caps were constructed and their assembly characterised.

3.1.1 System design

Our system is based on recombinantly expressing two fusion proteins, consisting of a pair of complementary oligomeric half-cage "caps". These use different intein systems to drive irreversible assembly. The half-cages are comprised of an oligomerisation domain, a rigid functionalisable domain, the intein driving force and affinity tags for purification (Figure 3.1). The oligomerisation domain specifies the geometry of the half cage by acting as the primary vertex, with the rigid functionalisable domain acting as the cage sides. All fusions are initially inert, and they only react when mixed with a compatible construct. When mixed, the intein driving force catalyses the half cage caps to fuse via spontaneous native chemical ligation to form the cage (Section 1.7).

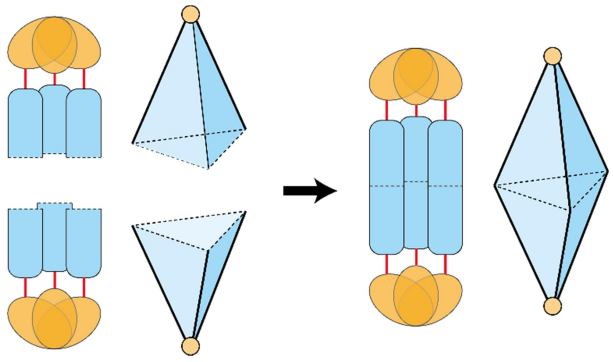


Figure 3.1 Schematic diagram of the designed complementary half cage caps with the vertices, β-trefoil knot homotrimer M4P (yellow ovals), the alpha-helical linkers that link the vertices and sides (red lines), and the side, CTPRs (blue rectangular). Upon mixing, the differing intein assembly systems join them together and form a trigonal bipyramidal caged product.

3.1.2 Trimeric half-cage caps

To validate our approach and explore the limits of our systems (structures formed, reaction speeds, efficiencies and yields), differing complementary trimeric half-cage caps were designed. The half cage caps were composed of: (i) the homotrimer Monofoil-4-P (M4P) domain as the primary vertex (Figure 3.2), (ii) consensus-designed tetratricopeptide repeat-containing protein (CTPR) as the sides (Figure 3.3), and (iii) differing intein assembly systems attached to their termini. An alpha-helical linker, $(EAAAK)_2$, connects the M4P and (CTPR) proteins, projecting the CTPR units away from each other and thereby reducing the risk of misfolding. The half-cage caps were engineered such that they would react to form a trigonal bipyramidal caged product Figure 3.1.

3.1.2.1 Homotrimeric vertex

The 48 amino-acid homotrimeric Monofoil-4-P (M4P) was selected as the vertex (Figure 3.2). Lee *et. al* designed the M4P by fragmenting the repeating primary structure of the fibroblast growth factor protein (J. Lee et al. 2011). They narrowed the domain that folds into a thermostable β-trefoil knot 42 residues domain, with a K_D of 10-50 nM. M4P was chosen due to its stability and the positions of its N and C termini. Both termini are exposed on the same plane and are not involved in the formation of the β-trefoil knot. Hence, extension via the addition of protein domains at either the N or C termini are able to produce complementary fusions without disrupting its homo-trimeric behaviour.

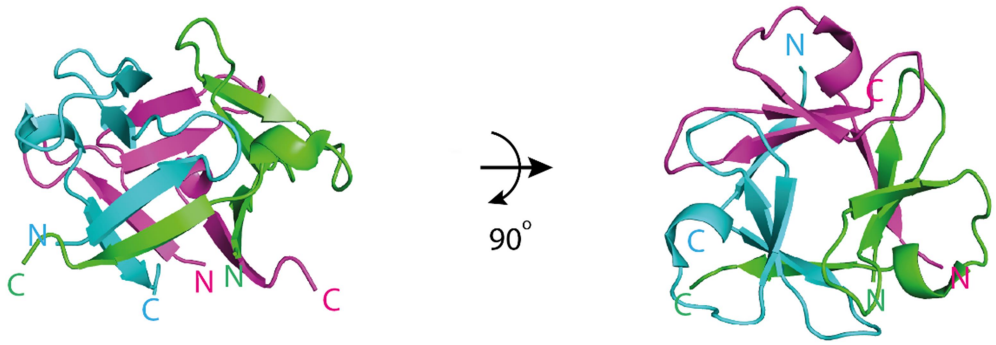


Figure 3.2 The designed homotrimer M4P chosen as the vertices for the structures. The side view and the top view of the crystal structure of the trimer are shown with each monomer coloured differently (magenta, green and turquoise). Each of the N and C termini is exposed on the same plane as labelled. The crystal structure of the M4P obtained from PDB 3OL0.

3.1.2.2 Rigid, rod-like sides

The rod-like consensus tetratricopeptide repeat proteins (CTPRn) were chosen to be the sides of the tripod half-cage caps due to their rigidity, stability and symmetry (Main et al. 2003; J J Phillips et al. 2012). Past studies have shown that these are monomeric, helix-turn-helix motif proteins that easily fold into rod-like structures. Most importantly, CTPRn can be extended from end to end, forming a superhelix (Main, Stott, et al. 2005; Jonathan J. Phillips, Millership, and Main 2012). In addition, they can be functionalised through the penta-peptide binding pocket located on the concave region of the CTPR (A. L. Cortajarena et al. 2004). For example, several CTPRs have been modified to contain binding pockets that bind to differing penta-peptide tags (Speltz, Nathan, and Regan 2015; Grove et al. 2012). In our study the CTPRs were produced without the C-terminal capping helix. This enables docking of the structures in a head to tail conformation. Thus, for example, when discussing CTPRs in the following chapters CTPR3 corresponds to 3 consensus TPRs only.

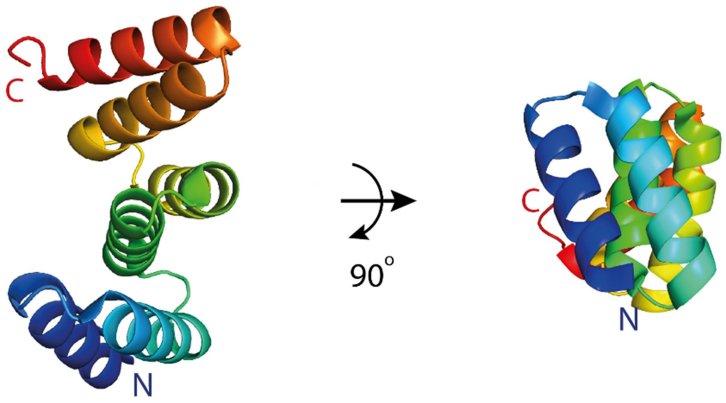


Figure 3.3 The sides of each half cage were composed of the repeat protein CTPR3. The crystal structure of CTPR3 is shown without its final C-capping solvating helix and with two helices per repeat. The crystal structure of the CTPR3 obtained from PDB 1NA0.

3.2 Closing mechanism – Previous Work

Previously in the Main laboratory, Dr. J. Wright successfully produced trimeric half-cage caps and investigated cage formation via cross linking with bis(sulfosuccinimidyl)suberate (BS3) crosslinker reagent and disulphide bond formation (Wright 2018). In both cases higher order structures were formed. However, most of the product consisted of the half-cage caps reacting intra-molecularly, rather than inter-molecularly. Thus, the yield of higher order structures was very low.

3.3 1st Generation cage closure system using the Mxe GyrA intein

In order to increase the yield of reaction, differing intein systems were trialled as directional driving forces. In nature, inteins are found in the middle of genes. Once expressed, they post-translationally modify the proteins they are in by excising themselves while ligating the flanking polypeptides regions together (as discussed in Section 1.7). The ligation is irreversible and forms a peptide bond [via native chemical ligation (NCL)].

Our first-generation system was based on the Mxe GyrA (MxGA) intein for the directional driving force. This stemmed from work already completed in the Main Laboratory that used a modified MxGA intein system for one-pot and step-wise protein-fibre assembly (Jonathan J. Phillips, Millership, and Main 2012; Harvey, Itzhaki, and Main 2018). This is explained in detail in the introduction (Section 1.7). Briefly - each used a number of fusion proteins that contain the protein to be ligated (POI) attached to either a C-terminal MxGA intein, a protease activate-able N-terminal cysteine or both. The N-terminal cysteine is protected with an affinity tag. When activation is required, the tag is cleaved with TEV protease to reveal a reaction ready cysteine. The MxGA intein is activated by adding MESNa reducing agent. This induces self-cleavage and produces an exposed thioester. Upon mixing of the activated components, the C-terminal thioester spontaneously reacts with the exposed N-terminal cysteine to produce a peptide bond. The reaction proceeds via a transthioesterification reaction, followed by an S-N acyl shift which leads to native peptide bond formation (Figure 3.4). Dr. J. Harvey demonstrated that natively folded proteins in mild conditions and at

concentrations of 50 μM yields 75 % ligation in 24 hrs (Harvey, Itzhaki, and Main 2018).

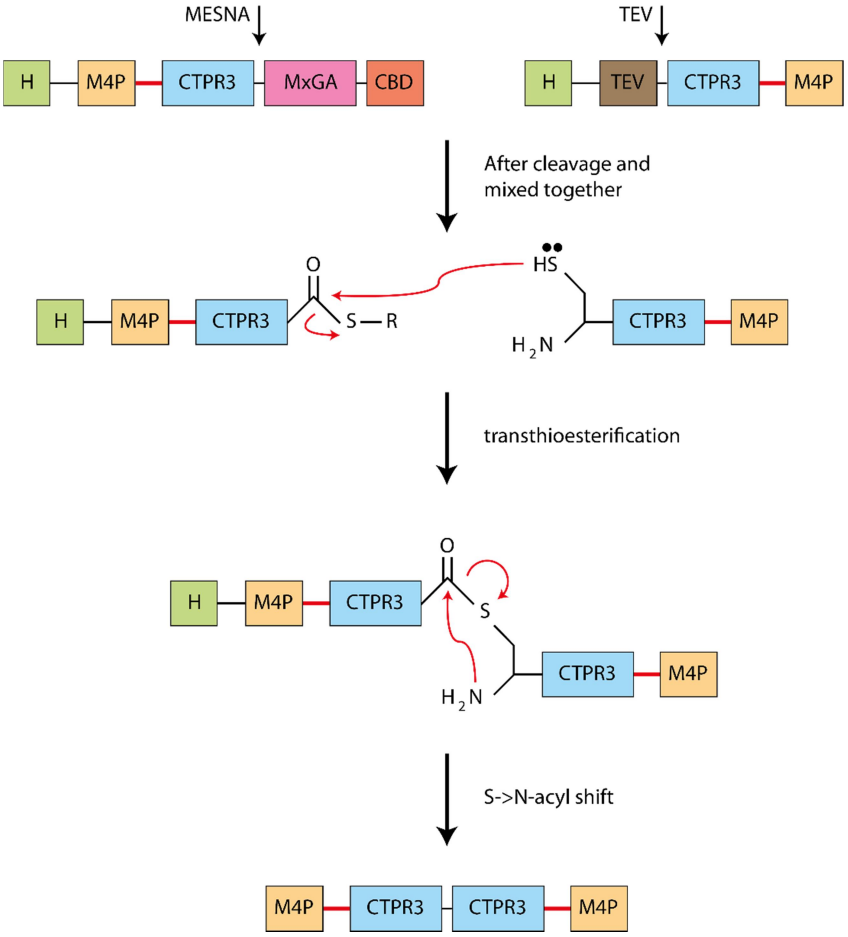


Figure 3.4 Schematic diagram of the NCL driven by intein, MxGA. The two half cage caps are activated prior to ligation. N-terminal half cage cap, H-M4P-CTPR3-MxGA-CBD is cleaved with MESNA to remove MxGA-CBD and expose a thioester group, while the C-terminal half cage cap, H-TEV-CTPR3-M4P is cleaved with TEV protease at the TEV cleavage site to remove H-TEV and reveal a cysteine. Upon mixing, transthioesterification took place leading to a S-N acyl shift and the irreversible formation of a native amide peptide bond.

3.3.1 Recombinant protein design

To adapt the system to direct cage formation two protein fusions were constructed as follows:

(1) H-M4P-CTPR3-MxGA-CBD – the MxGA intein and CBD were fused to the C-terminal of the half-cage caps and

(2) H-TEV-Cys-CTPR3-M4P – the cysteine is introduced to the N-terminus of the complementary half-cage cap.

As one would expect, both half-cage caps require activation prior to reaction. H-M4P-CTPR3-MxGA-CBD is activated by cleaving away the MxGA-CBD with MESNa, resulting in a C-terminal thioester (H-M4P-CTPR3-thio). Whereas, the activation of H-TEV-CTPR3-M4P is achieved by cleaving with TEV protease, exposing an N-terminal cysteine (cys-CTPR3-M4P). This enables two homotrimers half-cage caps to fuse together to form a single structure via NCL (Figure 3.5).

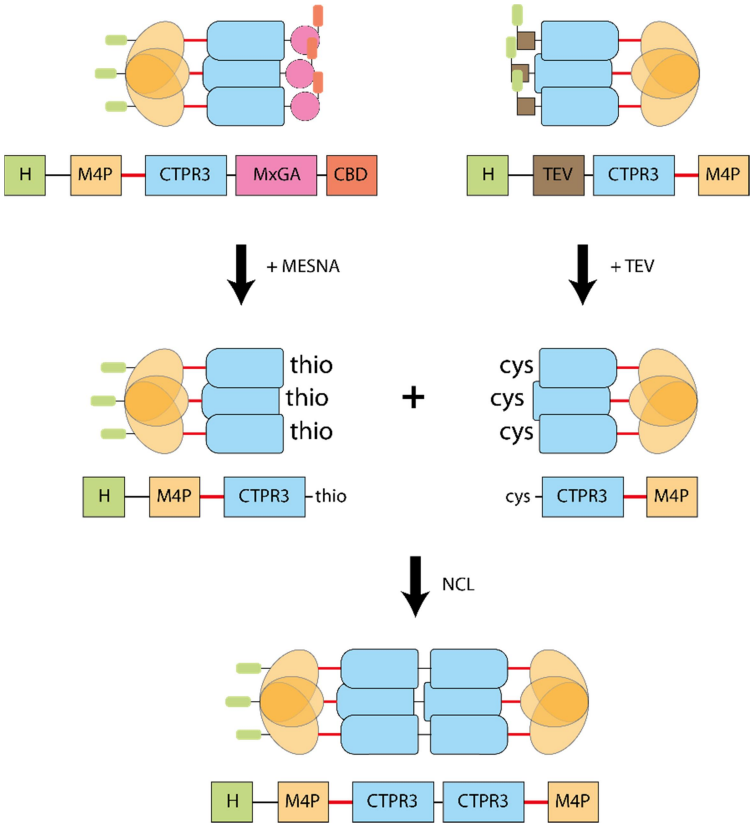


Figure 3.5 A schematic diagram where the two activated homotrimer half-cage caps fuse together to form a single structure via NCL. The cleavage of the MxGA-CBD was cleaved by MESNa during chitin affinity chromatography. Thus, the H-M4P-CTPR3-thio can be eluted after cleavage. The H-TEV-CTPR3-M4P is cleaved by TEV protease to reveal a cysteine. The activated half cage caps allow NCL to take place when mixed together. Green rectangle represents 6-Histidine tag; orange oval represents M4P; red line represents linker; blue rectangle represents CTPR3; pink circle represents MxGA; red rectangle represents CBD; and brown square represents TEV cleavage site.

3.3.2 Recombinant expression, purification and trimerisation analysis of H-M4P-CTPR3-thio and H-TEV-CTPR3-M4P

3.3.2.1 Expression and purification of H-M4P-CTPR3-thio and H-TEV-CTPR3-M4P

H-M4P-CTPR3-MxGA-CBD and H-TEV-CTPR3-M4P were successfully produced in high yields via recombinant expression and native purification (Section 2.3.3). Figure 3.6 shows the successful purification with yields of 8.4 mg/mL and 50 mg/mL, respectively. It is noteworthy that (i) H-M4P-CTPR3-MxGA-CBD was purified via chitin affinity chromatography (binds the CBD). Elution of pure reaction ready product occured via MESNa induced MxGA cleavage to give H-M4P-CTPR3-thio (i.e. C-terminal thioester). (ii) The observed sizes on the SDS-PAGE gel appear smaller than the expected. This is due to a gel-shift phenomenon caused by the CTPRs. Previous studies have shown that the CTPRs have a more compact structure under standard denaturing conditions (Aitziber L Cortajarena, Wang, and Regan 2010). These partially unfolded structures may affect the SDS binding and thus exhibit a size difference on the gel similar to other helix-turn-helix membrane spanning proteins (Rath et al. 2009).

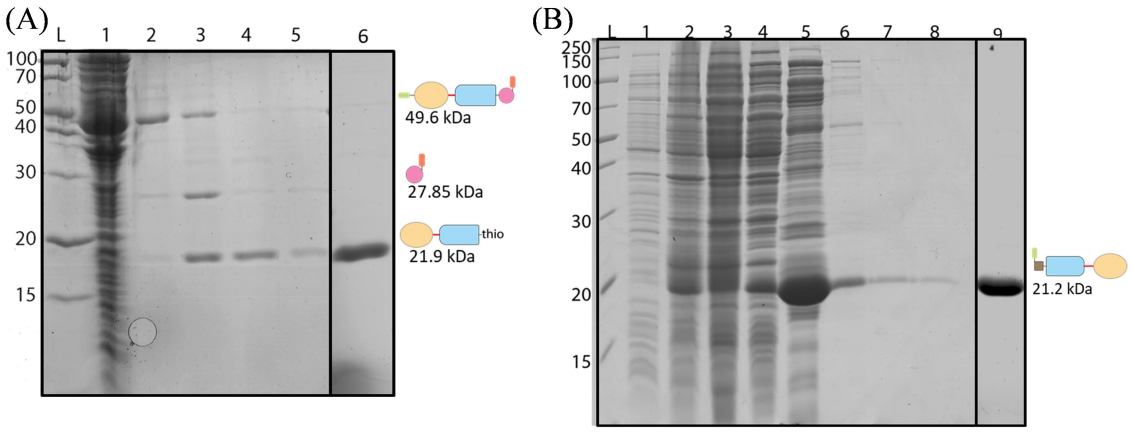


Figure 3.6 SDS-PAGE analysis of **(A)** the expression and purification of H-M4P-CTPR3-MxGA-CBD and C-terminal thioester production, H-M4P-CTPR3-thio, and **(B)** the expression and purification of H-TEV-CTPR3-M4P. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. **(A)** Lane 1, post-induction cell sample; Lane 2, post-wash chitin resin; Lane 3, 24 hrs cleavage with MESNa on chitin resin; Lane 4 and 5, chitin elution (H-M4P-CTPR3-thio); Lane 6, purified H-M4P-CTPR3-thio. **(B)** Lane 1, pre-induction cell sample; Lane 2, post-induction cell sample; Lane 3, native cell lysate; Lane 4, flow-through fraction; Lanes 5-8, elution fractions; Lane 9, purified H-TEV-CTPR3-M4P. The green rectangle represents 6-Histidine tag; orange oval represents M4P; red line represents linker; blue rectangle represents CTPR3; pink circle represents MxGA; red rectangle represents CBD; and brown square represents TEV cleaving site.

3.3.2.2 Trimerisation analysis of H-M4P-CTPR3-thio and H-TEV-CTPR3-M4P

The trimeric state of the H-M4P-CTPR3-thio and H-TEV-CTPR3-M4P was confirmed using SEC analysis in 50 mM Tris pH 8, 100 mM NaCl buffer (Figure 3.7). Fitting of the H-M4P-CTPR3-thio and H-TEV-CTPR3-M4P peak maxima to calibration standards, gave molecular weights of 73-76 kDa (calculated trimeric molecular weight: 65.7 kDa) and 63.6 kDa (calculated trimeric molecular weight: 63.6 kDa), respectively (Figure 3.7). The differences in expected and observed molecular weights are within 10 % error for this technique. SEC also confirms that trimerisation of the half cage caps are not concentration dependant in range used in this study. H-M4P-CTPR3-thio elutes as a single monodisperse peak at concentrations of 10, 50 and 100 µM (Figure 3.7).

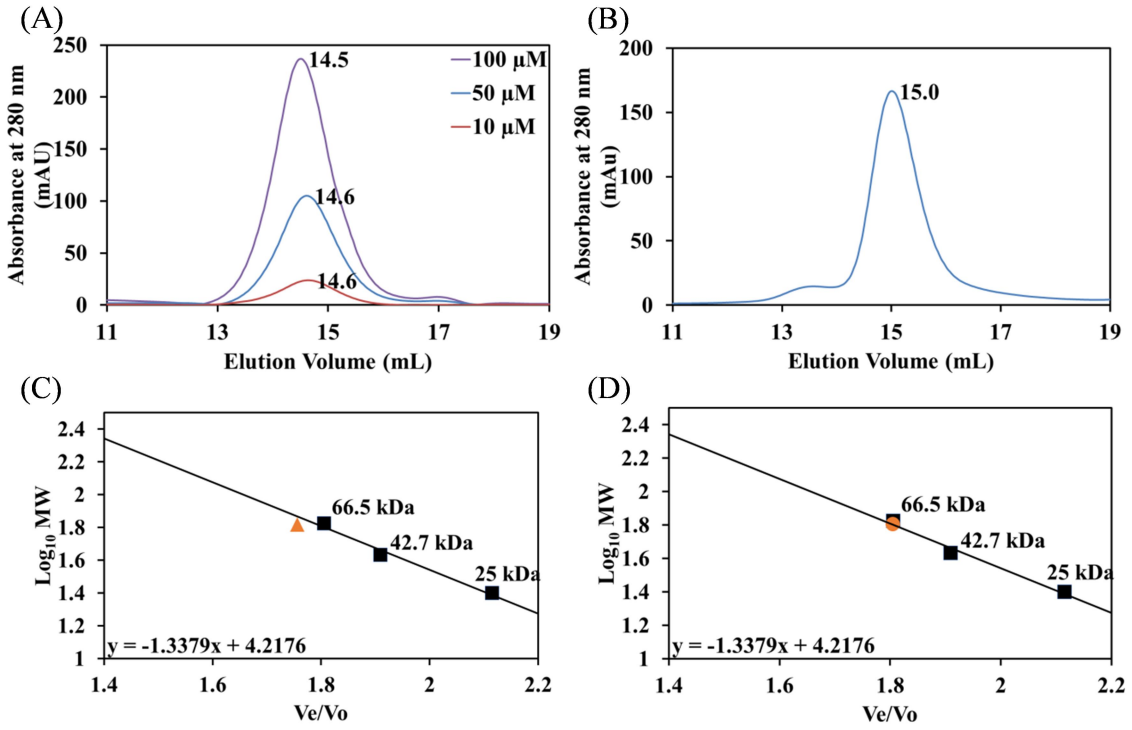


Figure 3.7 Superdex 200 10/30 SEC analysis of trimeric **(A)** H-M4P-CTPR3-thio and **(B)** H-TEV-CTPR3-M4P **(A)** Trimeric H-M4P-CTPR3-thio at different concentrations, 100 µM (purple), 50 µM (blue) and 10 µM (red). **(B)** Trimeric H-TEV-CTPR3-M4P at 50 µM. **(C and D)** The half cages Ve/Vo (elution volume/column void volume) plotted against their Log_{10} molecular weights on a standard curve. The black squares are protein standards **(C)** The elution volume used was 14.6 mL. The orange triangle represents H-M4P-CTPR3-thio. **(D)** The orange circle represents H-TEV-CTPR3-M4P.

3.3.2.3 Activation of cys-CTPR3-M4P

H-TEV-CTPR3-M4P was activated to its N-cysteinyl form, cys-CTPR3-M4P, by TEV protease cleavage (10 U per mg of H-TEV-CTPR3-M4P substrate) in 50 mM Tris, 150 mM NaCl, 5 mM TCEP pH 8 for 16 hrs at 25 °C. Cleavage was estimated to reach almost 100 % by SDS-PAGE analysis. Post-cleavage cys-CTPR3-M4P was purified by nickel affinity chromatography to remove non-cleaved protein and increase purity (Figure 3.8).

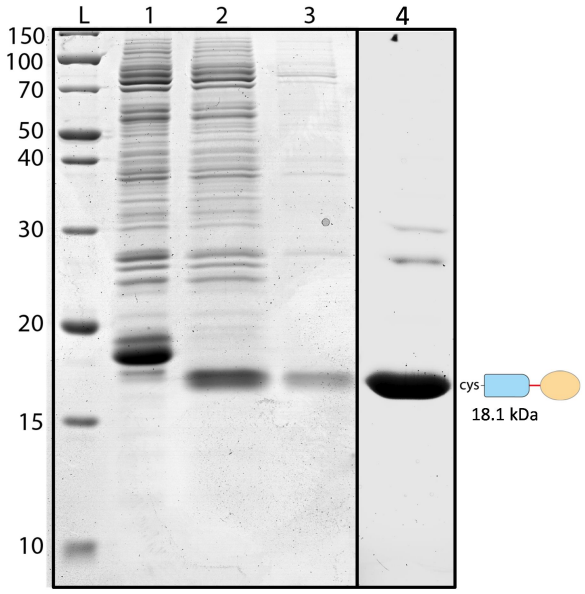


Figure 3.8 Activation of cys-CTPR3-M4P (18.1 kDa) by cleaving H-TEV-CTPR3-M4P (21.2 kDa) with TEV. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, pre-cleavage of H-TEV-CTPR3-M4P; Lane 2, post-cleavage of H-TEV-CTPR3-M4P; Lane 3, flow-through fraction from post-cleavage purification; and Lane 4, purified activated cys-CTPR3-M4P. The orange oval represents M4P; red line represents linker; and blue rectangle represents CTPR3.

3.3.3 NCL reaction of H-M4P-CTPR3-thio and cys-CTPR3-M4P

The two activated trimeric H-M4P-CTPR3-thio and cys-CTPR3-M4P were reacted in 50 mM Tris pH 8.5, 1 M NaCl, 200 mM MESNa, 10 mM TCEP for 24 hrs at 30 °C with a final protein concentration of 50 μM in 1:1 molar ratio of protein. These conditions have been shown to give high yielding reaction when used for fibre assembly (Harvey, Itzhaki, and Main 2018). SDS-PAGE analysis of the reaction is shown in Figure 3.9. Unfortunately, there are no significant signs of successful ligation between the two trimeric fusion proteins.

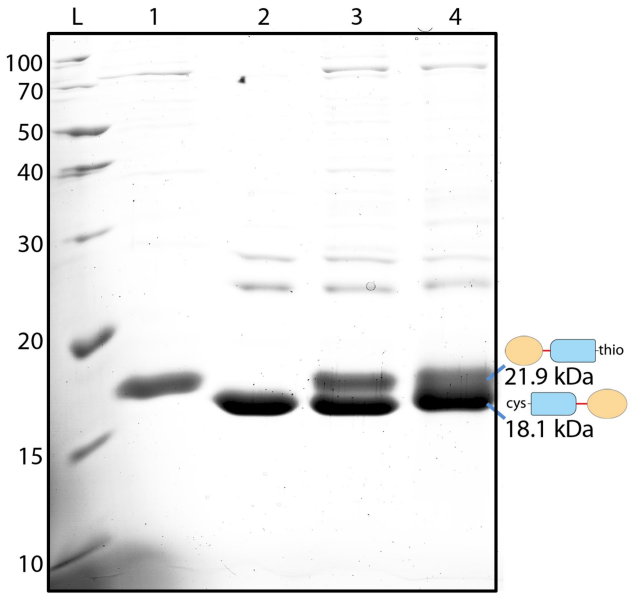


Figure 3.9 NCL reaction of the activated H-M4P-CTPR3-thio and cys-CTPR3-M4P. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, activated H-M4P-CTPR3-thio; Lane 2, activated cys-CTPR3-M4P; Lane 3, NCL reaction at time 0; Lane 4, NCL reaction at time 24 hrs. Orange oval represents M4P; red line represents linker; and blue rectangle represents CTPR3.

3.3.4 Summary

Both SEC and MALDI-TOF analysis confirm that the H-M4P-CTPR3-thio and cys-CTPR3-M4P are trimeric. However, the NCL reaction of the two half-cage caps failed. This work was repeated by Dr. J. Harvey whom obtained the same outcome. The failure of the NCL reaction is likely due to steric hindrance. Thus, Dr. J. Harvey also reacted the two half-cages in denaturant (4 M GuHCl and 4 M Urea). It was hoped that once denatured, the steric hindrance would decrease, enabling the reaction to proceed. However, no detectable ligation product was observed (Harvey 2016). Interestingly, it has been shown that the trimeric M4P is incredibly stable to both temperature and denaturant. For example, the Main laboratory has observed the trimeric half-cages by SDS-PAGE after boiling and in the presence of SDS. Thus, the reaction might require the half-cages to be left in 8 M GuHCl for 24 hrs prior to the reaction. Nevertheless, such a scheme would be too time consuming and remove many of the benefits of the system. Thus, it was decided to re-engineering the assembly system to use a differing driving force: Split-inteins.

3.4 2nd Generation cage closure system using the Split-inteins

Split-inteins are natural variants of intein domains that are divided into two separate polypeptide chains (Wu, Hu, and Liu 1998; Martin, Xu, and Evans 2001; Dassa et al. 2009; Zettler, Schütz, and Mootz 2009; Carvajal-Vallejos et al. 2012). Peptide chains with each split-intein half are inert until complementary halves are combined, where upon they spontaneously fold together to produce an active intein. The now active intein self-catalyses their excision and ligates the two separate peptide chains together (Figure 3.10).

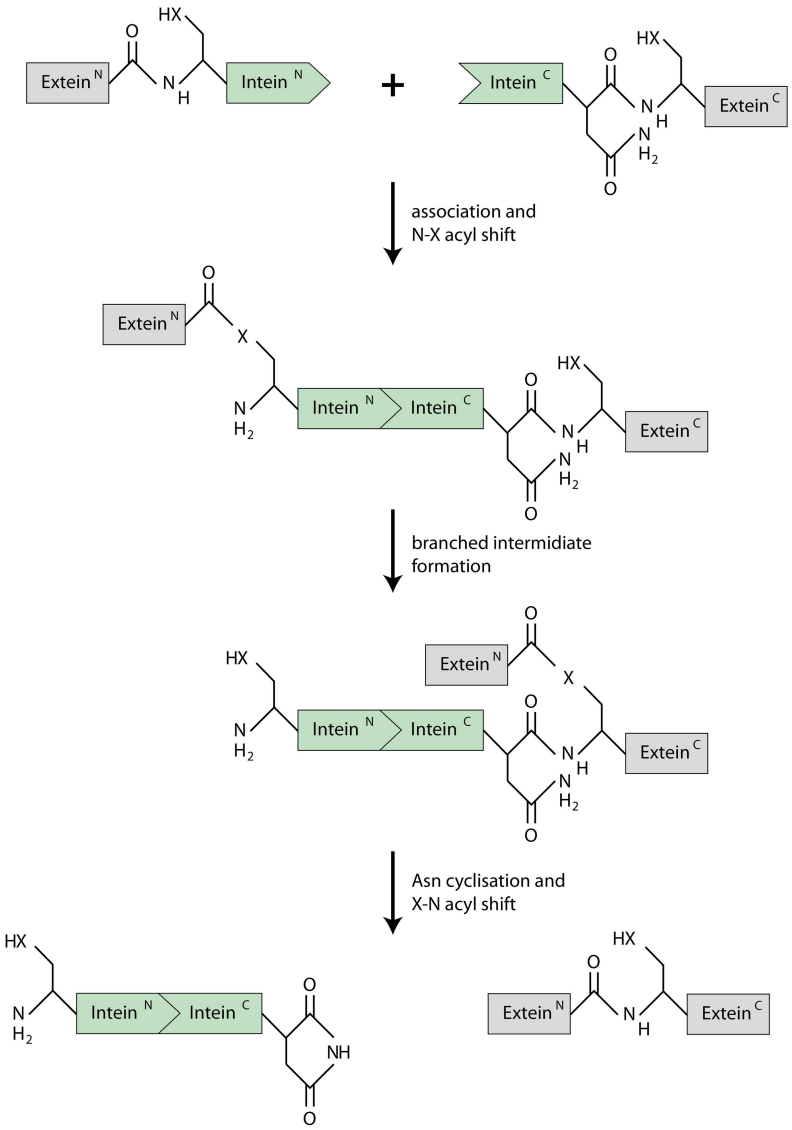


Figure 3.10 Schematic diagram of NCL driven by split-inteins. Following split-inteins association, an N–S or N–O acyl shift forms a thioester or oxoester bond at the N-extein/intein junction. This reactive intermediate is attacked in transthioesterification by the side chain sulfhydryl or hydroxyl group of the first residue in the C-extein, which can be Cys, Ser, or Thr, to give a branched intermediate. The cyclisation of the conserved Asn residue at the C-terminus of the intein releases the intein. Finally, the thioester bond between the exteins rearranges to a peptide bond by a spontaneous S–N or O–N acyl shift. X can be sulphur or oxygen (Carvajal-Vallejos et al. 2012).

3.4.1.1 Recombinant proteins design

In order to obtain a protein system that would enable step-wise assembly the Main laboratory (and Dr. J. Wright in particular) selected two natural split-intein pairs which were orthogonal but still produced fast and high yielding ligation. These were : Gp41-1 (Gp), a split-intein from gp41 DNA helicase and IMPDH-1 (Imp), a split-intein from inosine-5'-monophosphate dehydrogenase (Dassa et al. 2009). Both of these split-inteins have been shown to have high reaction rates and yields (90% in 10 mins), and have no cross-reactivity at 5 μM concentrations (Carvajal-Vallejos et al. 2012). However, the reaction does leave a short insertion of approximately 10 amino acid residues. This method has many advantages over the MxGA system. For example, it does not require pre-reaction activation and should be substantially faster and higher yielding (intra versus inter-molecular reaction).

Initially 2 sets of constructs were designed:

(1) H-M4P-CTPR3-Gp^N and (2) H-Gp^C-CTPR3-M4P,

(3) H-M4P-CTPR3-Imp^N and (4) H-GST-Imp^C-CTPR3-M4P.

In each case the N-terminus of the split-intein was fused to the C-terminus of the M4P-CTPR3 and the C-terminus of the split-intein was fused to the N-terminus of the CTPR3-M4P (Figure 3.11). Dr. J. Wright investigated the yield of Gp split-inteins while I focused on Imp split-inteins.

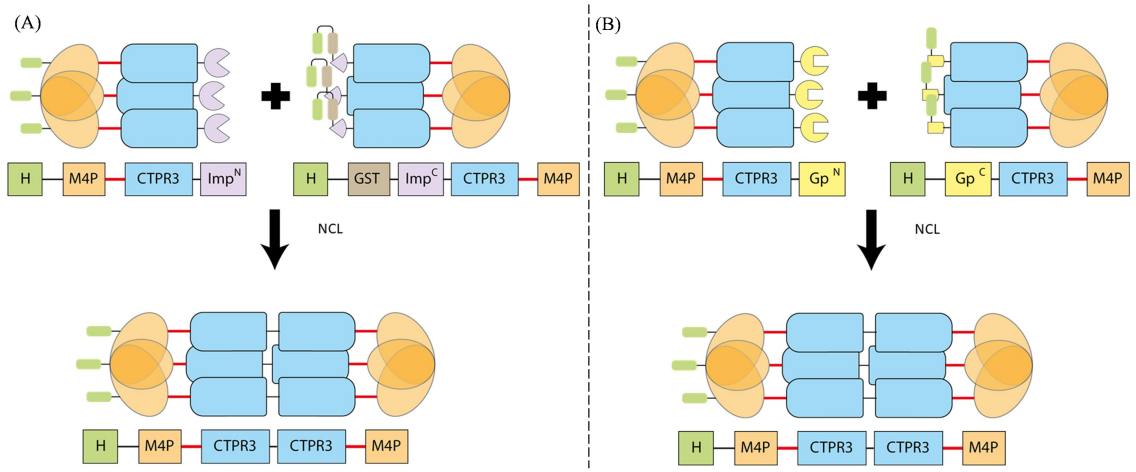


Figure 3.11 Schematic diagram of the formation of the trigonal bipyramidal cages via split-inteins NCL. Here, the N-terminal split-intein (**(A)** Imp^N and **(B)** Gp^N) is fused to the C-terminus of the N-terminal half-cage caps; and the C-terminal split-intein (**(A)** Imp^C and **(B)** Gp^C) is fused to the N-terminus of the C-terminal half-cage caps. Upon mixing, the split-inetins fold into an active enzyme that leads to NCL, joining the half-cage caps together while excising itself. Green represents the 6-Histidine tag; orange represents M4P; red represents the linker; blue represents CTPR3; purple represents Imp split-intein; brown represents CBD; and yellow represents Gp split-intein.

3.4.2 Recombinant expression, purification and trimerisation analysis of H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P

3.4.2.1 Recombination expression and purification of H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P

Initially both H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P were expressed recombinantly at differing temperatures and times. Unfortunately, no combination of expression conditions produced protein that could be purified natively. Therefore, a denaturing protocol was used (Section 2.3.3) and shown in Figure 3.12. This produced good yields (19.5 mg/mL and 39.5 mg/mL respectively) with high purity. However as both were eluted from the Ni affinity column in 8 M Urea, they required refolding. Both were dialysed overnight into reaction buffer (50 mM Tris pH 8, 300 mM NaCl, 5 mM DTT) with differing urea concentrations. No visible precipitation was observed when H-M4P-CTPR3-Imp^{N} was dialysed into reaction buffer with 1 M urea in it. In contrast H-GST-Imp^{C}-CTPR3-M4P showed no precipitation in 2 M and a small amount in 1M urea.

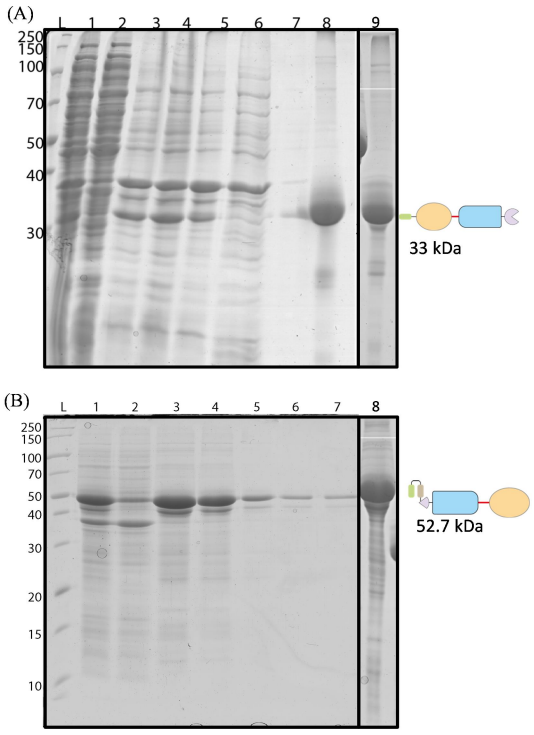


Figure 3.12 SDA-PAGE analysis of the purification of **(A)** H-M4P-CTPR3-Imp^{N} and **(B)** H-GST-Imp^{C}-CTPR3-M4P. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. **(A)** Lane 1, post-induction; Lane 2, native supernatant; Lane 3, native pellet; Lane 4, denatured supernatant; Lane 5, denatured pellet; Lane 6, flow-through fraction; Lane 7, wash fraction; Lane 8, elution fraction; and Lane 9, purified H-M4P-CTPR3-Imp^{N}. **(B)** Lane 1, post-induction denatured soluble cell lysate; Lane 2, flow-through fraction; Lane 3-7, elution fractions; and Lane 8, purified H-GST-Imp^{C}-CTPR3-M4P. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; and brown represents CBD.

3.4.2.2 Analysis of the refolded H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P

To determine if the refolding of H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P by dialysis was successful, analytical SEC was performed in their reaction buffers (Figure 3.13A and B). For H-M4P-CTPR3-Imp^{N}, a single symmetric peak at 13 mL was observed. For H-GST-Imp^{C}-CTPR3-M4P, the 1 M urea refolded sample eluted in the void volume. However, the dialysed sample in 2 M urea eluted with a peak at 12.5 mL.

Comparison of the peak maxima for H-M4P-CTPR3-Imp^{N} obtained 1 M urea and H-GST-Imp^{C}-CTPR3-M4P obtained in 2 M urea with to calibration standards, gave molecular weights of 145 kDa (32 % larger than the calculated trimeric molecular weight of 99 kDa) and 160 kDa (1 % larger the calculated mass of 158.1 kDa for a homotrimeric species) respectively (Figure 3.13C and D). Thus, the H-M4P-CTPR3-Imp^{N} half cage caps run slightly larger than the protein standards predict for its trimeric molecular weight. This suggests that the H-M4P-CTPR3-Imp^{N} is a less spherical structure than H-GST-Imp^{C}-CTPR3-M4P.

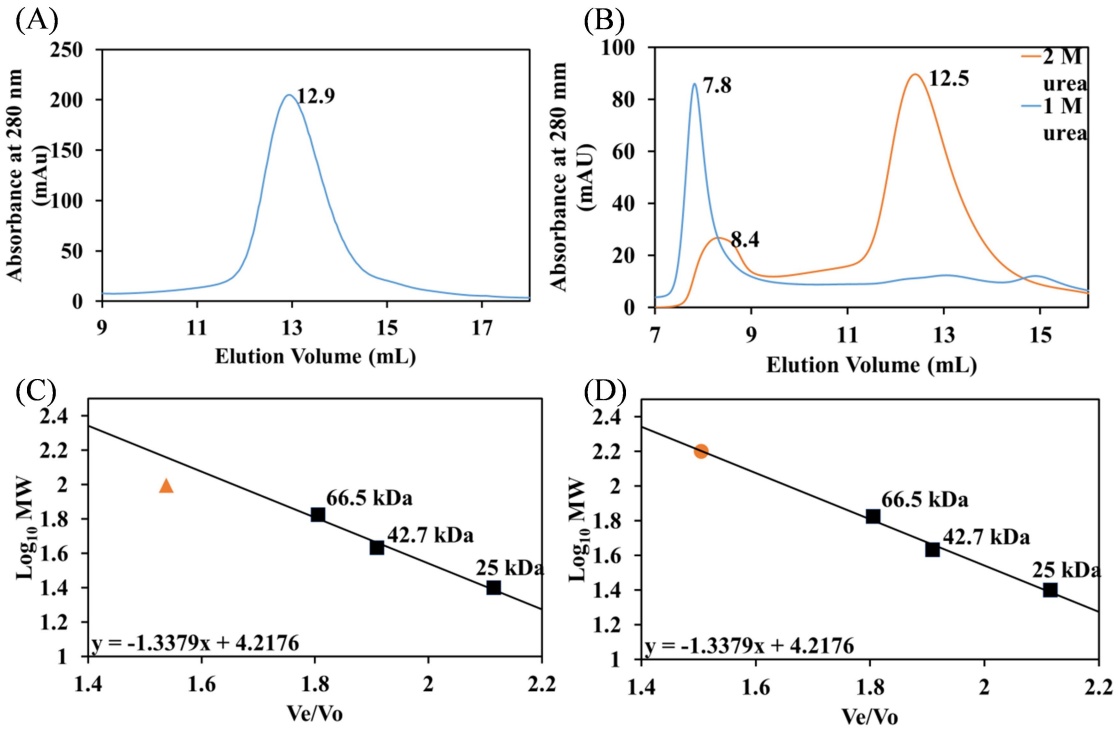


Figure 3.13 Superdex 200 10/30 SEC analysis of refolded trimeric **(A)** H-M4P-CTPR3-Imp^{N} and **(B)** H-GST-Imp^{C}-CTPR3-M4P. Trimeric H-GST-Imp^{C}-CTPR3-M4P in 1 M urea (blue) and 2 M urea (orange). **(C and D)** The half cages Ve/Vo (elution volume/column void volume) plotted against their Log_{10} molecular weights on a standard curve. The black squares are protein standards. **(C)** The orange triangle represents H-M4P-CTPR3-Imp^{N}. **(D)** The orange circle represents H-GST-Imp^{C}-CTPR3-M4P.

3.4.3 [1]Recombinant expression, purification and trimerisation analysis of H-M4P-CTPR3-Gp^N and H-Gp^C-CTPR3-M4P

Both H-M4P-CTPR3-Gp^N and H-Gp^C-CTPR3-M4P expressed well and were purified successfully in denaturing conditions with high yield and purity (12 mg/mL and 25 mg/mL respectively). The protein was refolded whilst bound to nickel resin to 0 M urea (Figure 3.14). Comparison of the peak maxima for H-M4P-CTPR3-Gp^N and H-Gp^C-CTPR3-M4P obtained in 0 M urea with calibration standards, gave molecular weights of 95.1 kDa (calculated trimeric molecular weight: 106 kDa) and 79.8 kDa (calculated trimeric molecular weight: 91 kDa), respectively (Figure 3.15) (Wright 2018). The profile shows that the half cage caps run slightly larger than the protein standards predict for their trimeric molecular weights. This suggests that they are non-spherical.

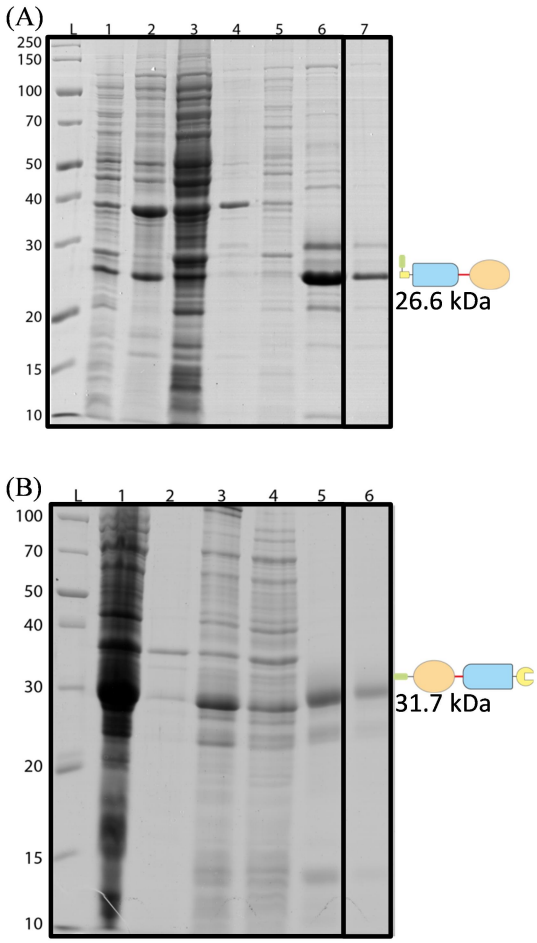


Figure 3.14 SDS-PAGE analysis of the purification of **(A)** H-M4P-CTPR-Gp^N and **(B)** H-Gp^C-CTPR-M4P (Data obtained by Dr. J Wright). **(A)** Lane 1, post-induction; Lane 2, insoluble lysate; Lane 3, soluble lysate; Lane 4, flow-through fractions; Lane 5, wash step fractions; Lane 6, elution fraction; and Lane 7, purified H-M4P-CTPR-Gp^N. **(B)** Lane 1, post-induction; Lane 2, insoluble lysate; Lane 3, soluble lysate; Lane 4, flow-through fraction; Lane 5, elution fraction; and Lane 6, purified H-Gp^C-CTPR-M4P. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; and yellow represents Gp split-intein.

[1] The recombinant expression and purification of H-M4P-CTPR3-Gp^N and H-Gp^C-CTPR3-M4P were performed by Dr. J. Wright.

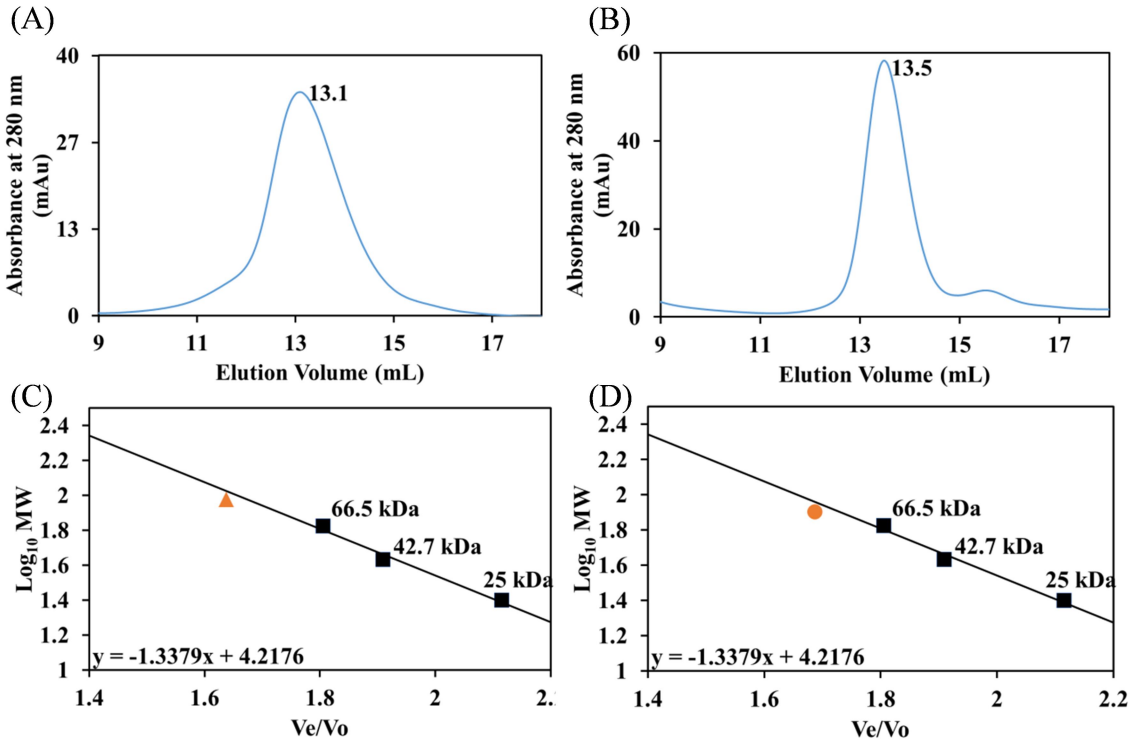


Figure 3.15 Superdex 200 10/30 SEC analysis of refolded trimeric **(A)** H-M4P-CTPR3-gp^{N} and **(B)** H-Gp^{C}-CTPR3-M4P (Data obtained from Dr. J. Wright). **(C and D)** The half cages Ve/Vo (elution volume/column void volume) plotted against their Log_{10} molecular weights on a standard curve. The black squares represent protein standards. **(C)** The orange triangle represents H-M4P-CTPR3-Gp^{N}. **(D)** The orange circle represents H-Gp^{C}-CTPR3-M4P.

3.4.4 NCL reaction of 2nd generation cage closure system[1]

Ligation reactions of both Imp and Gp cages were initiated by mixing purified cognate half-cages in equimolar concentration of 50 µM under mild conditions as described in Section 2.4.2. The reaction buffer was supplemented with 1 M urea for the Gp reactions (to avoid any aggregation) and 2 M urea for the Imp mediated ligations due to the solubility of H-GST-Imp^{C}-CTPR3-M4P. Samples were taken at several time points over a 24 hr period and analysed by SDS-PAGE (Figure 3.16). Excitingly, both reactions successfully produced significant yields of ligated product. The product H-M4P-CTPR3-CTPR3-M4P (40.6 kDa) was observed in 5 mins (Imp split-intein mediated ligation) and 1 min (Gp split-intein mediated ligation) at ~32 kDa on both SDS-PAGE gels. Comparing the ligation reactions of both split-inteins, we found that the ligation reaction mediated by Imp split-intein took significantly longer to complete and with a much lower yield *i.e.* 50 % yield after 24 hrs as opposed to 85 % yield in 30 mins. This is most likely due to the increased urea concentration used in the reaction buffer and the presence of partially folded GST tag.

[1] The ligation reactions mediated by Gp split-inteins were performed by Dr. J. Wright.

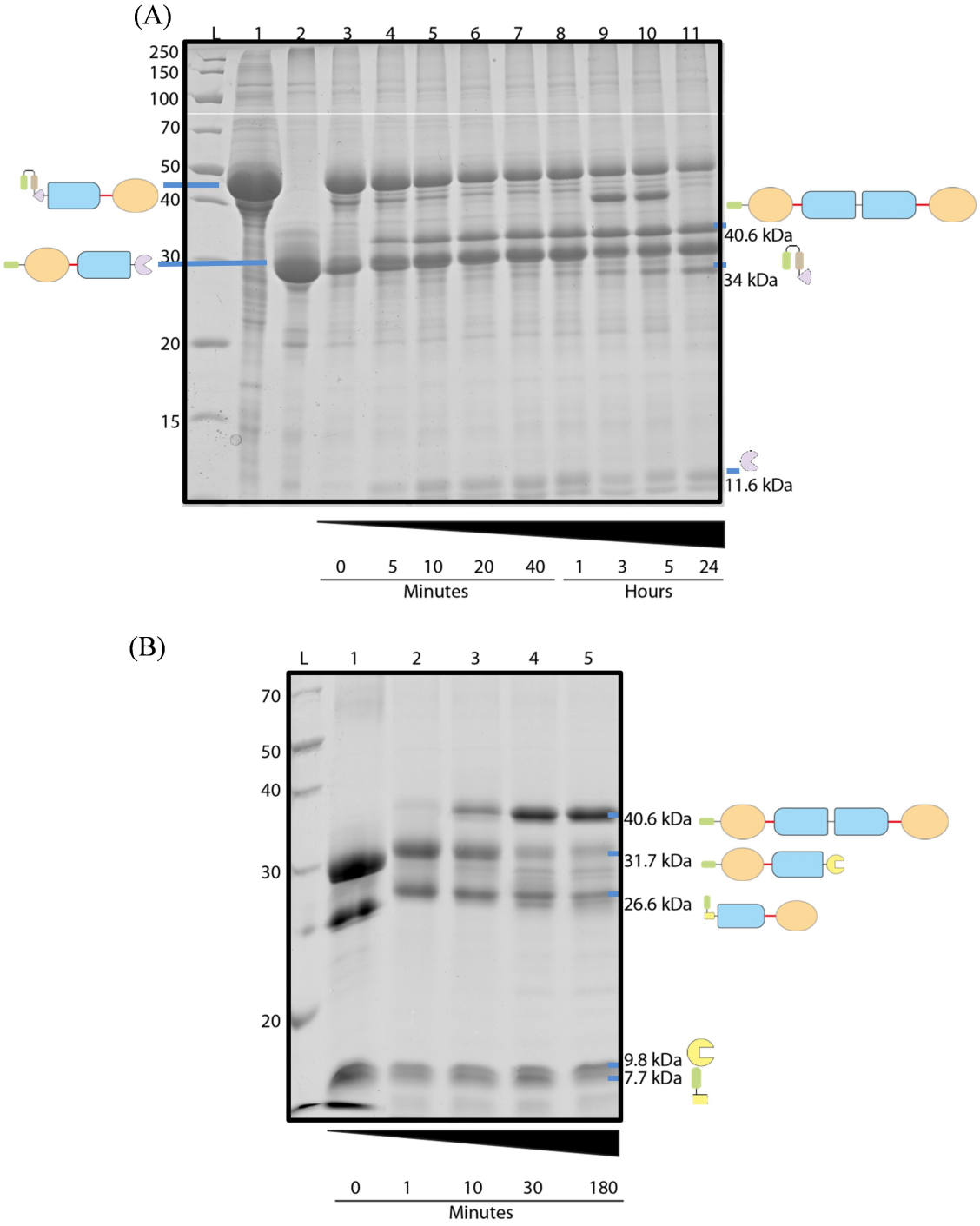


Figure 3.16 SDS-PAGE analysis of the reaction between **(A)** H-M4P-CTPR3-Imp^{N} and H-GST-Imp^{C}-CTPR3-M4P over 24 hrs, and **(B)** H-M4P-CTPR3-Gp^{N} and H-Gp^{C}-CTPR3-M4P over 3 hrs (data collected by Dr. J. Wright). **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; **(A)** Lane 1, H-GST-Imp^{C}-CTPR3-M4P; Lane 2, H-M4P-CTPR3-Imp^{N}; Lane 3, 0 mins; Lane 4, 5 mins; Lane 5, 10 mins; Lane 6, 20 mins; Lane 7, 40 mins; Lane 8, 1 hr; Lane 9, 3 hrs; Lane 10, 5 hrs; and Lane 11, 24 hrs. An extra band can be observed at time points 3 and 5 hrs (Lanes 9 and 10); this may due to insufficient boiling of the samples, causing the product to be in a dimeric/trimeric state. **(B)** Lane 1, 0 mins; Lane 2, 1 min; Lane 3, 10 mins; Lane 4, 30 mins; and Lane 5, 180 mins. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; brown represents CBD; and yellow represents Gp split-intein.

3.4.5 Purification of ligated products

The characterisation of the ligation product requires the removal of the unligated reactants and excised split-inteins. Unfortunately, in the 2nd Gen designs, all reactants and products possess similar affinity tags. Thus, affinity chromatography could not be used to separate any of the ligation mixture. Therefore, SEC was trialled for the Gp cage closure reactions (Figure 3.17). It was decided to focus on the Gp closure as ligation, rather than those mediated by the Imp split-intein, due to the far greater yields. High yields are extremely important in any reaction and particularly so here, given that each trimeric half cage requires each of its three sides to react. For example, if the yield of ligated product is lower than 67 % it indicates that, for every trimeric half-cage cap, one "arm" has not reacted.

The preparative SEC of the Gp split-intein ligation mixture gave a broad peak. Unfortunately, when the major peak was analysed by SDS-PAGE, it was found that both ligated and unligated half-cage caps eluted together (Figure 3.17). Thus, although the excised split-inteins can be removed by SEC, the ligated product and unligated reactants cannot.

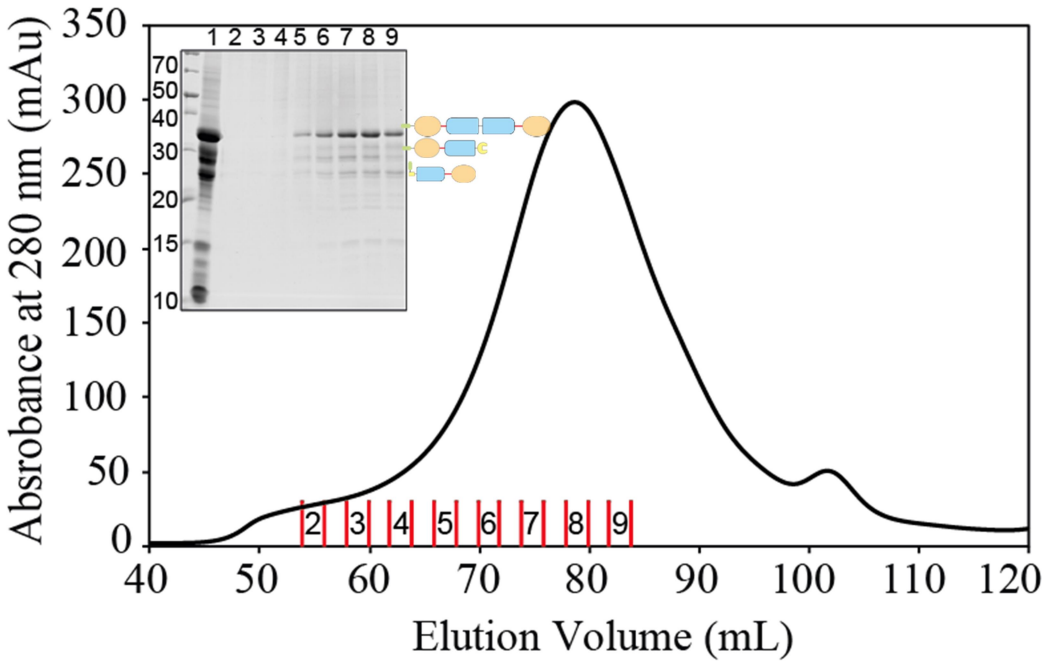


Figure 3.17 Preparative SEC of a 50 µM Gp mediated cage ligation reaction after 24 hrs in 1 M urea with the denaturing SDS-PAGE Gel of the major peak. 5 mL of the post ligation reaction mixture was injected onto a Superdex^{TM} S200 16/60 preparative column running standard ligation buffer. Top right corner: Lanes of the SDS-PAGE gel correspond to the fractions labelled in the graph. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; and yellow represents Gp split-intein.

3.4.6 Summary

All four half cage caps with split-inteins were expressed and successfully refolded into their trimeric states in high yields. However, H-GST-ImpC-CTPR3-M4P did require at least 2 M urea to refold into its trimeric form. Importantly, all half cage cages could be successfully ligated. The Gp mediated assembly produced a better yield in a shorter time compared to Imp mediated assembly (85 % yield in 5 hrs and 50 % yield in 24 hrs, respectively). This is most probably due to the higher concentration of denaturant in the Imp reaction and the presence of partially refolded GST tag. Significantly, the fusion protein design and, more specifically, the placement of the affinity tags within each fusion protein, exposed a major limitation - we were unable to separate ligated product from unreacted half-cage caps. Therefore, the system was re-designed to enable easier purification of the ligated product.

3.5 3rd Generation Cage closure system using the Split-inteins

To separate the fully ligated product from partially ligated and unreacted half-cage caps, the fusion constructs were changed to enable affinity purification. The change was relative straightforward and involved moving the His-affinity tag on the N-terminus of the H-M4P-CTPR3-Gp^N and H-M4P-CTPR3-Imp^N to the C-terminus: M4P-CTPR3-Gp^N-H and M4P-CTPR3-Imp^N-H. Thus, upon ligation, all affinity tags will be excised along with the split-inteins. As a result, there is no affinity tag on the ligated product. When subjected to affinity chromatography, the ligated product will not bind, enabling easy purification. In addition, the H-GST-Imp^C-CTPR3-M4P construct was further modified to remove the GST-tag to aid refolding. In its place a CBD tag was added to aid solubility. This is because the Imp^C split-intein is known to require an additional domain close by to aid production (Personal Communication, J. Wright). The final construct was: H-CBD-Imp^C-CTPR3-M4P. Finally, a CBD tag was also added to the M4P-CTPR3-Imp^N-H construct: M4P-CTPR3-Imp^N-CBD-H. This was to ensure that, when ligated, the cage product would not overlap with the reactants on SDS-PAGE. Figure 3.18 shows the improved constructs reaction scheme for both split-intein systems.

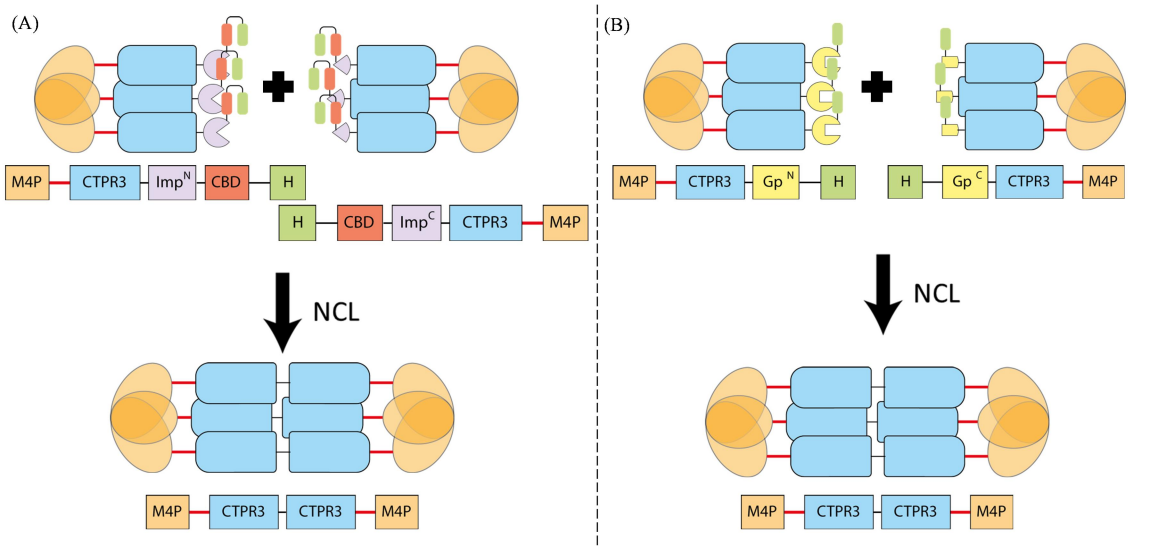


Figure 3.18 Schematic diagram of the formation of the trigonal bipyramidal cages via 2nd generation split-inteins NCL system. Here, all affinity tags are fused to split-inteins. Upon mixing, the split-inetins fold into an active enzyme which leads to NCL, joining the half-cage caps together while excising itself along with the affinity tags. The completely ligated product does not contain any affinity tag and therefore can be purified via affinity chromatography. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split- intein.

3.5.1 Recombinant expression, purification and trimerisation analysis of M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P

3.5.1.1 Recombination expression and purification of M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P

M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P were successfully expressed and purified as per Section 2.3.3with high yields and purity (20.5 mg/mL and 15.0 mg/mL respectively) (Figure 3.19). M4P-CTPR3-Imp^{N}-CBD-H was purified natively. In contrast, the H-CBD-Imp^{C}-CTPR3-M4P could only be purified via a denaturing purification. H-CBD-Imp^{C}-CTPR3-M4P was easily refolded either by dialysis or whilst being purified via affinity chromatography into reaction buffer supplemented with 1 M urea.

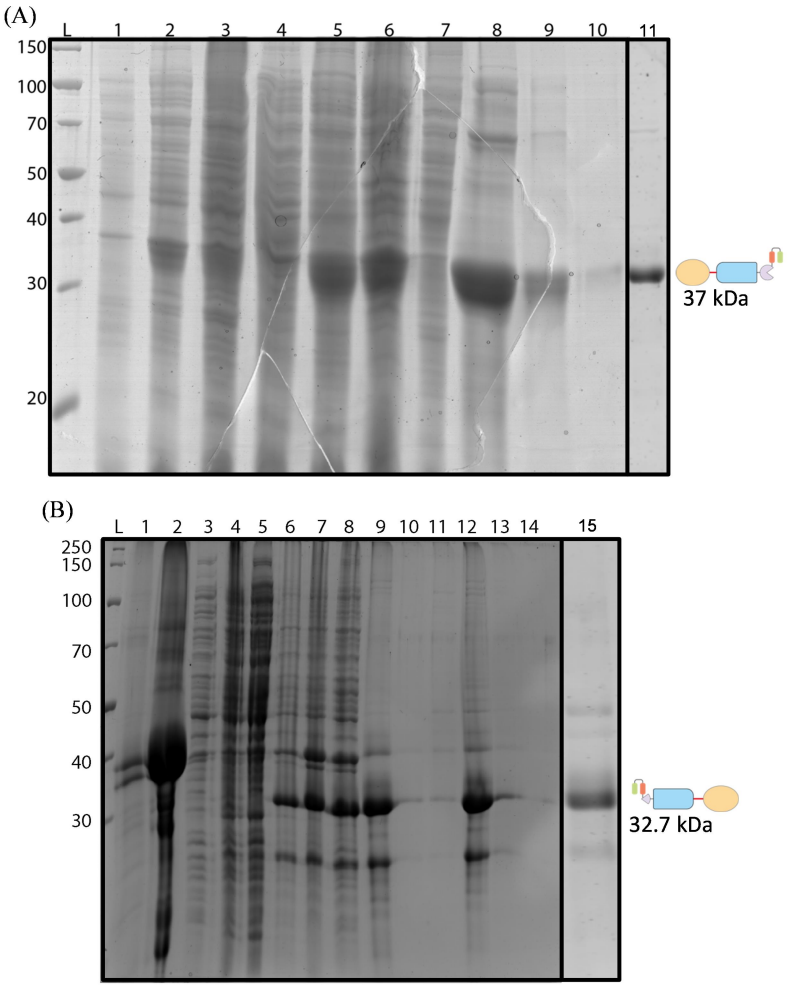


Figure 3.19 SDA-PAGE analysis of the purification of **(A)** M4P-CTPR3-Imp^{N}-CBD-H and **(B)** H-CBD-Imp^{C}-CTPR3-M4P. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, pre-induction; Lane 2, post-induction; Lane 3, native supernatant. **(A)** Lane 4, native pellet; Lane 5, denatured supernatant; Lane 6, denatured pellet; Lane 7, flow-through fraction; Lane 8-10, elution fractions; and Lane 11, purified M4P-CTPR3-Imp^{N}-CBD-H. **(B)** Lane 4, denatured supernatant; Lane 5, denatured pellet; Lane 6, flow-through fraction; Lane 7-14 elution fractions; and Lane 15, purified H-CBD-Imp^{C}-CTPR3-M4P. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; and red represents CBD.

3.5.1.2 Trimerisation analysis of M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P

The trimeric states of the M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P were confirmed using SEC analysis in reaction buffer with 1 M urea (Figure 3.20). Fitting of the M4P-CTPR3-Imp^{N}-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P peak maxima to calibration standards, gave 139 kDa (calculated trimeric weight of 111 kDa) and 97.8 kDa (calculated trimeric weight of 98.1 kDa), respectively (Figure 3.20). This suggests that the M4P-CTPR3-Imp^{N}-CBD-H is a more elongated shape than H-CBD-Imp^{C}-CTPR3-M4P.

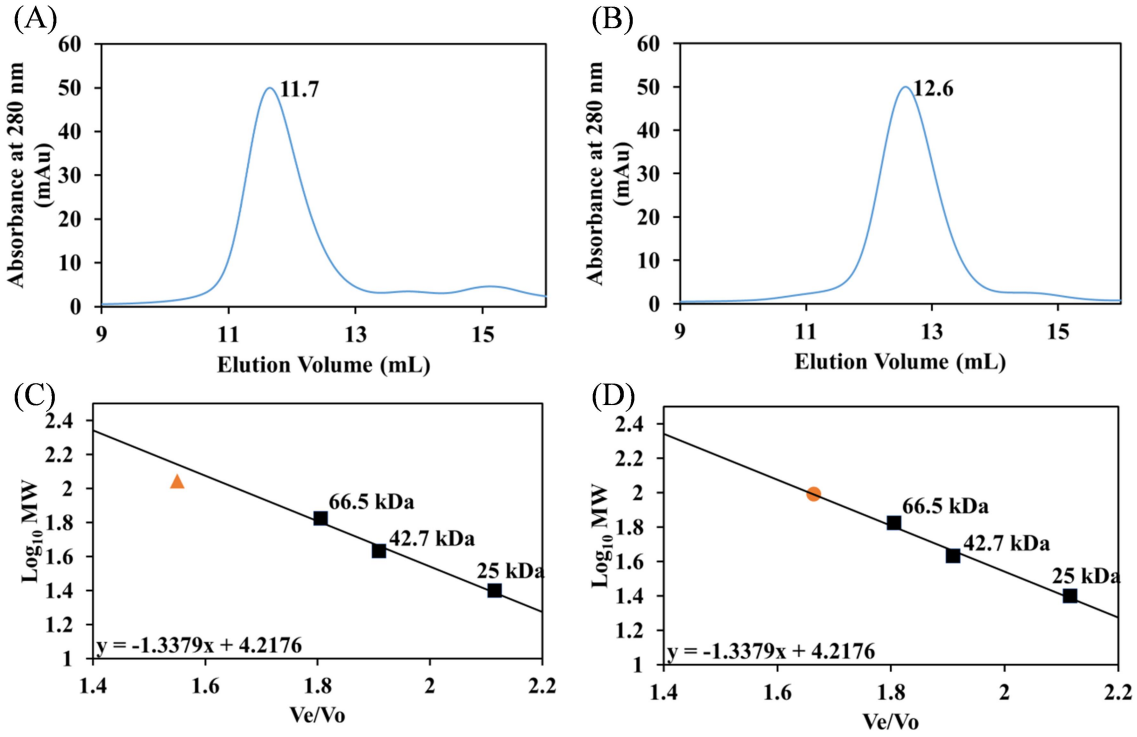


Figure 3.20 Superdex 200 10/30 SEC analysis of refolded trimeric **(A)** M4P-CTPR3-Imp^{N}-CBD-H and **(B)** H-CBD-Imp^{C}-CTPR3-M4P in reaction buffer with 1 M urea. **(C and D)** The half cages Ve/Vo (elution volume/column void volume) plotted against their Log_{10} molecular weights on a standard curve. The black squares are protein standards. **(C)** The orange triangle represents M4P-CTPR3-Imp^{N}-CBD-H. **(D)** The orange circle represents H-CBD-Imp^{C}-CTPR3-M4P.

3.5.2 [1]Recombinant expression, purification and trimerisation analysis of M4P-CTPR3-Gp^N- H

M4P-CTPR3-Gp^N-H was successfully expressed and purified under denaturing conditions (10 mg/L - Figure 3.21). The recombinant protein was dialysed into reaction buffer containing 1 M urea with little precipitation. SEC confirmed that the recombinant protein was successfully refolded into its trimeric state (Figure 3.22). Fitting of the peak maxima for M4P-CTPR3-Gp^N-H to calibration standards, gave 129.4 kDa [calculated trimeric molecular weight (94.1 kDa)].

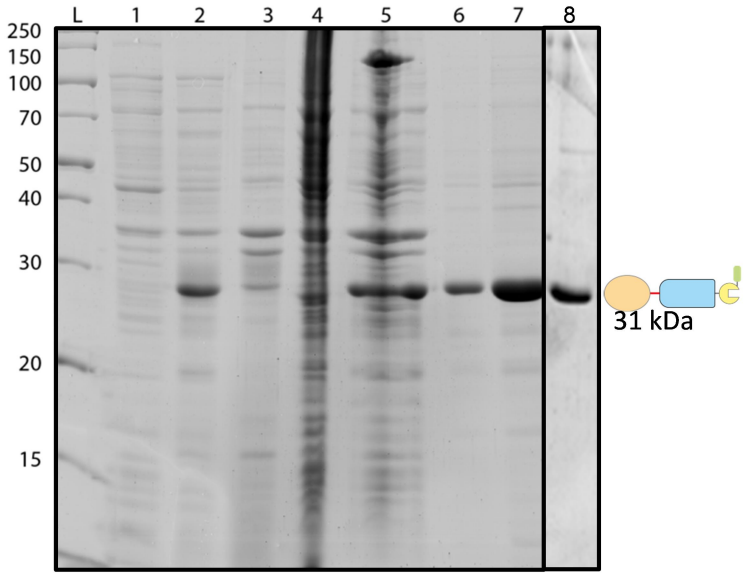


Figure 3.21 SDS-PAGE gels analysis of the purification of M4P-CTPR-Gp^N-H. Lane 1, pre-induction culture; Lane 2, post-induction culture; Lane 3, denatured insoluble lysate; Lane 4, denatured soluble lysate; Lane 5, flow-through fraction; Lane 6-7, elution fractions and Lane 8, purified M4P-CTPR-Gp^N-H (Data obtained by Dr. J Wright). Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; and yellow represents Gp split-intein.

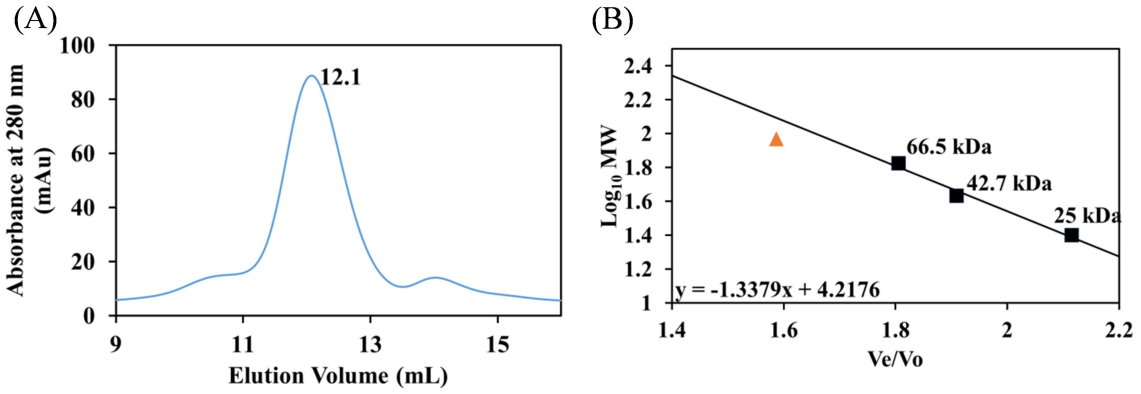


Figure 3.22 **(A)** Superdex 200 10/30 SEC analysis of refolded trimeric M4P-CTPR3-Gp^N-H (Data obtained by Dr. J. Wright). **(B)** The half cages Ve/Vo (elution volume/column void volume) plotted against their Log_{10} molecular weights on a standard curve. The black squares represent protein standards; and orange triangle represents M4P-CTPR3-Gp^N-H.

[1] All ligation reactions mediated by Gp split-inteins were performed by Dr. J. Wright.

3.5.3 NCL reaction of 3rd generation cage closure system[1]

Ligation reactions were initiated by mixing purified cognate half-cages in equimolar concentrations from 1 μM to 200 μM under mild conditions (reaction buffer with 1 M Urea – Section 2.4.2). Samples were taken at several time points over a 24 hr period and analysed by SDS-PAGE (Figure 3.23). As with the 2nd generation cages, both Gp and Imp mediated ligation reactions produced successful ligations at all concentrations. Figure 3.23 shows a typical reaction conducted at the 100 μM, for both Gp and Imp mediated ligations. Importantly, both Imp and Gp produced rapid, high yielding ligations. Interestingly, even though Gp mediated reactions have a higher reaction rate than Imp, they gave lower overall yields (Figure 3.24). Both split-intein-mediated ligations produced ≥ 65 % yield within 10 mins, and the Imp reactions reaching ≥ 80 % and Gp ≥ 70 % within 3 hrs. After 3 hrs, all the reactions were close to completion with only small additional increases in yield when the reactions were left for 24 hrs. Moreover, there was very little difference in the final yields across the protein concentration range of 1-100 μM for either Imp or Gp ligations. At the higher concentration of 200 μM, the reactions produced some protein precipitation leading to a small reduction in yields. Interestingly, although the yields were very high, we did not achieve the ~95 % values of the previous study (Carvajal-Vallejos et al. 2012). This is consistent with the theory that the structure of the tripod half-cage caps creates steric hindrance that reduces the efficiency. In the case of NCL by MxGA intein (Section 3.3), the trimeric structures stopped any reaction; whereas here, the steric hindrance reduced the splicing efficiency by 10 to 30 % (depending on the split-intein and reaction conditions used).

[1] All ligation reactions mediated by Gp split-inteins were performed by Dr. J. Wright.

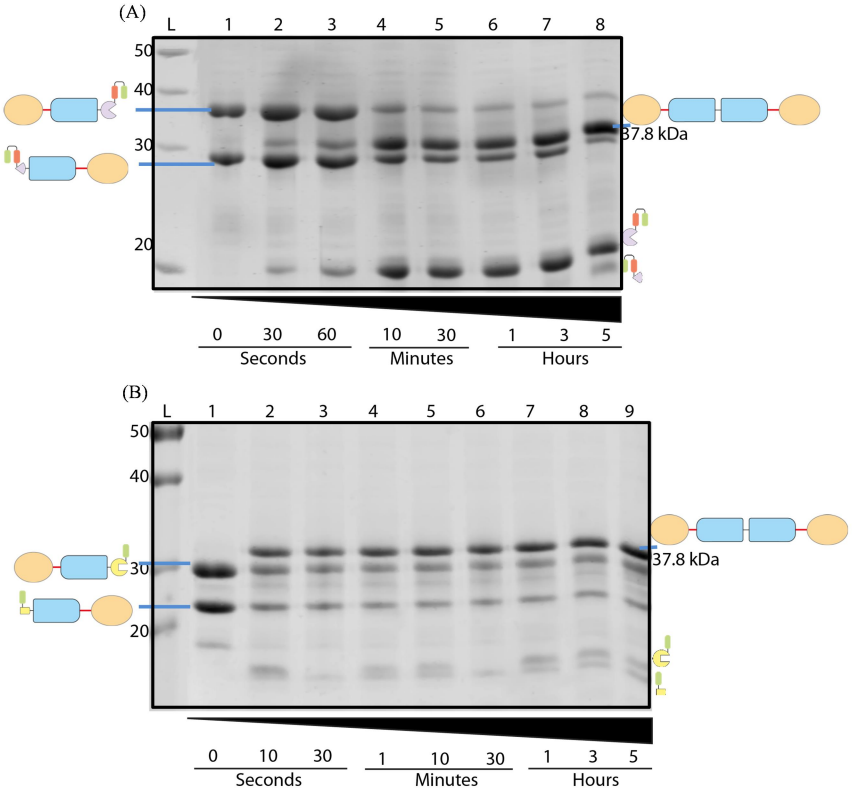


Figure 3.23 Example of the SDS-PAGE of NCL 3rd generation split-intein reaction in 100 μM final concentration. **(A)** Reaction of Imp split-inteins: M4P-CTPR3-ImpN-CBD-H and H-CBD-Imp^{C}-CTPR3-M4P; **(B)** reaction of Gp split-inteins: M4P-CTPR3-Gp^{N}-H and H-Gp^{C}-CTPR3-M4P. Data collected by Dr. J. Wright. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, 0 sec; and Lanes 2-9, various time points. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

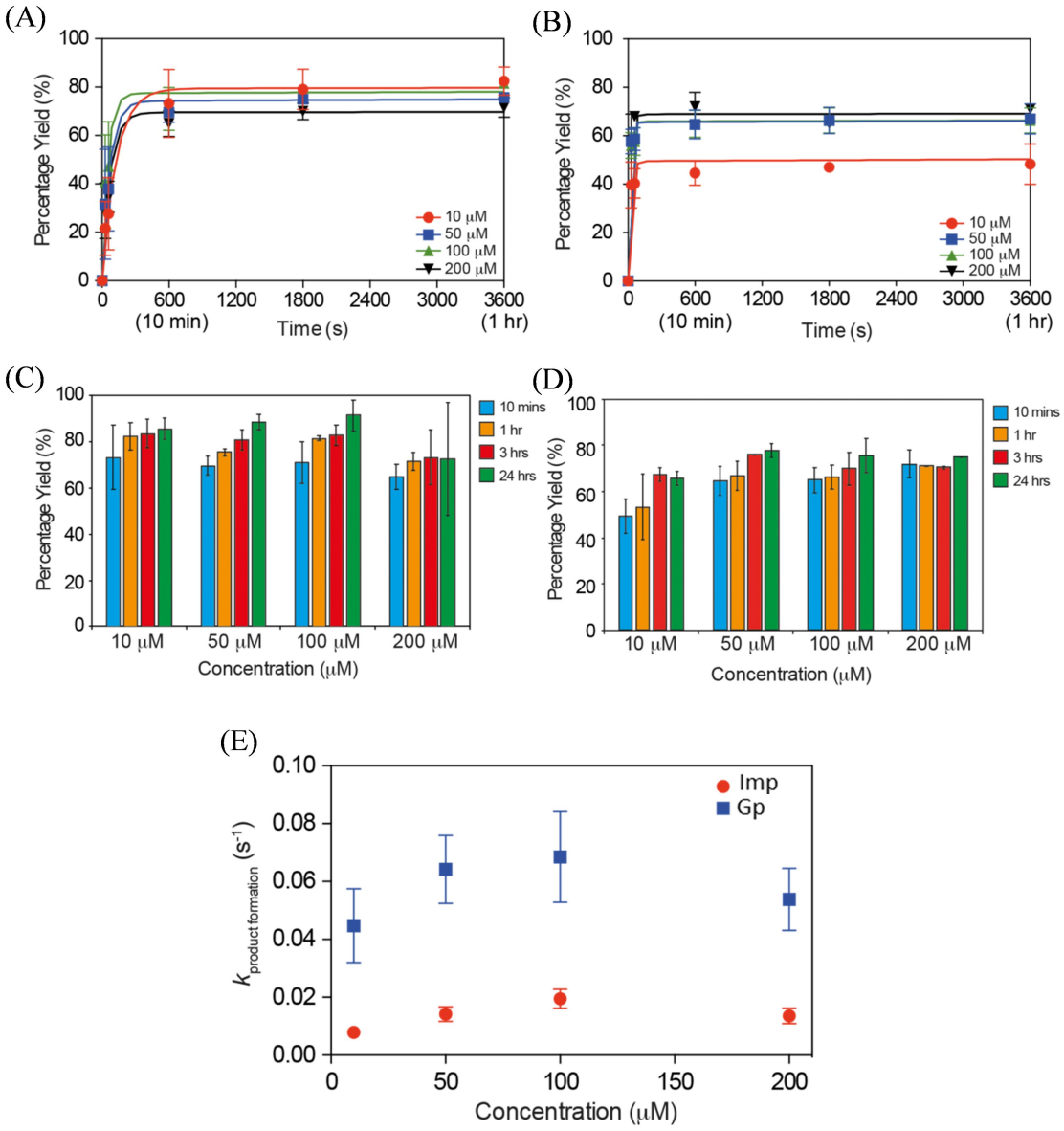


Figure 3.24 **(A and B)** Graphs of the initial rates of percentage production formation obtained from SDS-PAGE fitted with a single exponential plus linear drift [(A*(1 -exp(k*t))) + (m*t)] **(A)** 2nd generation Imp split-inteins mediated NCL and **(B)** 2nd generation Gp split-inteins mediated NCL. **(C and D)** Graphs of the percentage production formation up to 24 hrs after the reaction was initiated. Error bars equate to standard deviation of multiple repeat experiments (at least three in all cases). **(C)** 2nd generation Imp split-inteins mediated NCL and **(D)** 2nd generation Gp split-inteins mediated NCL; and **(E)** Comparison of rate constants of ligation production formation for Imp and Gp mediated NCL at differing protein concentrations. Error bars equate to standard error of fit to initial rates.

3.5.4 Purification of completely ligated products

To separate the fully ligated assemblies from unreacted and excised split-inteins, nickel affinity chromatography was performed. After the first affinity chromatography step, all ligation reactions contained > 95 % purity of fully ligated assemblies as assayed by SDS-PAGE and anti-His affinity tag Western blot (Figure 3.25).

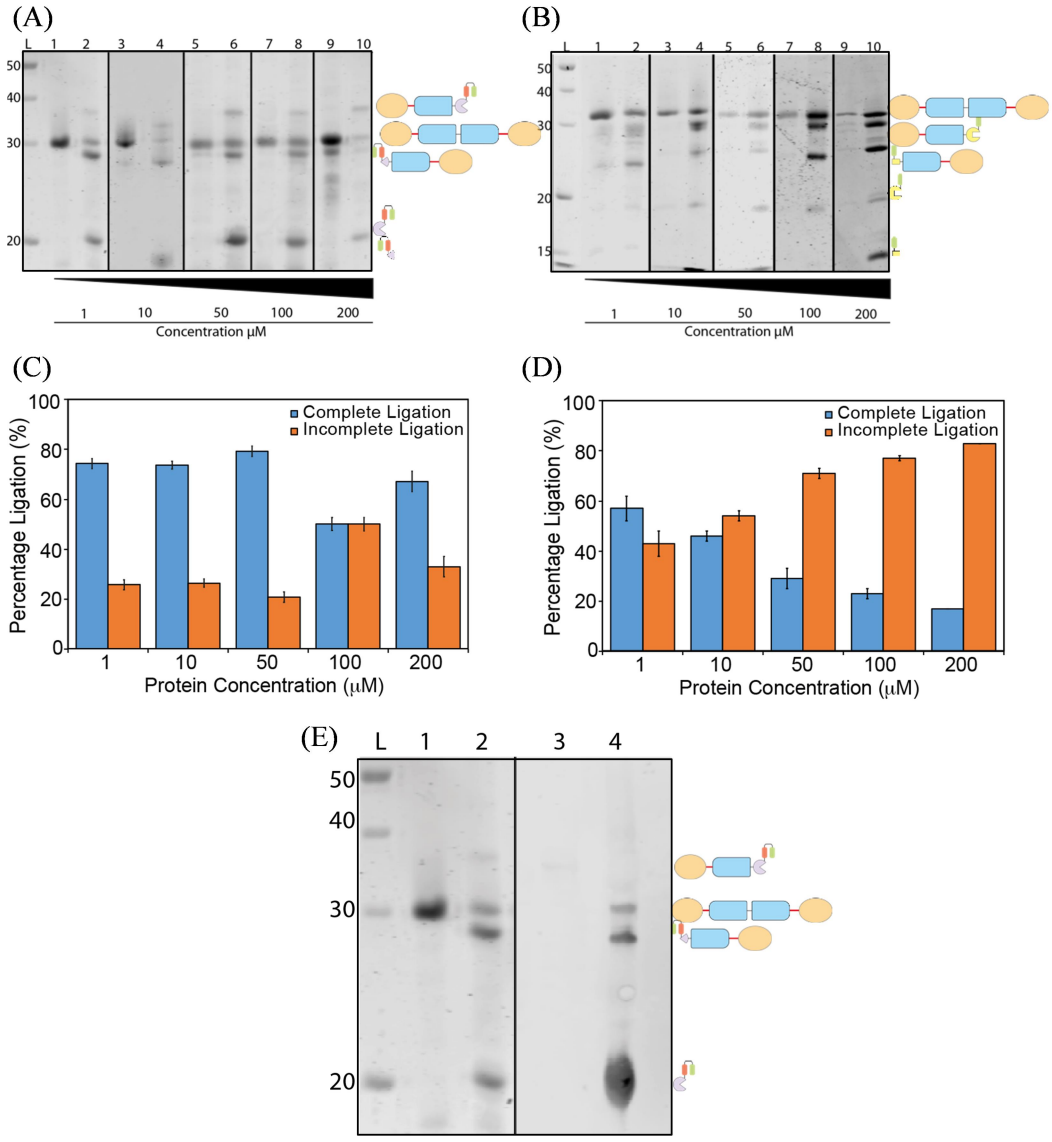


Figure 3.25 SDS-PAGE gel analysis of reaction purification at a range of concentrations. **(A)** Imp split-intein system. **(B)** Gp split-intein system (Data collected by Dr. J. Wright). **(A and B)** Odd number lanes are flow-through and wash fractions, even numbers are elution fractions: Lanes 1 and 2, 1 µM reaction concentration; Lanes 3 and 4, 10 µM reaction concentration; Lanes 5 and 6, 50 µM reaction concentration; Lanes 7 and 8, 100 µM reaction concentration; Lanes 9 and 10, 200 µM reaction concentration. **(C, D)** Calculated percentage of the complete ligation and partial ligation for the Imp and Gp split-intein systems respectively, at a range of concentrations Error bars equate to standard deviation of multiple repeat experiments (at least three in all cases). **(E)** SDS-PAGE and anti-His Western blot of Imp split-inteins system in 1 µM reaction, aligned on border, Lanes 1 and 2, SDS-PAGE; Lanes 3 and 4, Western blot: Lanes 1 and 3, flow-through and wash fractions; Lanes 2 and 4, elution fractions. **(A, B and E)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue rectangle represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

Interestingly, SDS-PAGE showed that there is "ligated product" bound to the Ni affinity column during the purification step. This stems from partially ligated trimeric structures. *i.e.* where at least one 'arm' of the trimeric half-cage cap has been ligated, but where the others were not. It is interesting to speculate that the partially ligated products may stem from the irreversibility of the reaction. For example, when A-B-C subunits from one cap react with a-b-c from another, they can link in a correct manner giving a discrete cage (A-a, B-b and C-c). However, they can also react to produce partially ligated faulty "dead-end" structures (A-a, B-c).

The relative intensities of the bands in the flow-through and elution samples were used to estimate the percentage of partially and completely ligated products. Importantly, the amount of completely ligated product increased as the concentration of reactants is lowered. From the analysis, it is clear that the Gp mediated ligations lead to more partial ligation and less cage closure than the Imp mediated ligations. For example, Imp mediated ligations at lower protein concentrations (1 μM to 50 μM) all produced > 75 % fully ligated product. In contrast, Gp mediated ligations at lower protein concentrations (1 μM and 10 μM) produced only ~ 50 % fully ligated product. Thus, the faster ligation speed of the Gp split-inteins hinders discrete cage formation and leads to the production of higher proportions of partially ligated structures and networks.

3.5.5 Analysis of ligated product and purification of discrete cages

The purified ligation reactions were analysed using size-exclusion chromatography (SEC) (Figure 3.26). This analysis showed that all ligations generated, to a greater or lesser extent, heterogeneous mixture of differently sized proteins. That is, in addition to the expected fully ligated discrete cages, the ligation reactions also produced networks of "cross-ligated" proteins. In both the Imp and Gp split-inteins mediated NCL, higher ligation concentrations exhibited both a larger void volume peak and broader major peak. Excitingly, and in contrast, lower protein concentrations, *i.e.* < 50 μM, gave a better separated monodisperse peak at 12 mL elution volume. Overall, the ligation mediated by Imp split-inteins produce a more monodisperse peak compared to Gp. This elution peak when fitting to calibration standards gave an estimated mass of 122 kDa, consistent with the calculated mass of a homotrimeric cage product (113 kDa) (Figure 3.26C).

Thus, to obtain high quantities of purified discrete caged product the Imp mediated ligation reactions were conducted at 10 µM overnight in reaction buffer supplemented with 1 M urea. They were purified using the two-stage process outlined above (Ni affinity chromatography followed by SEC). This produced a high purity, homogeneous sample for further structural analysis.

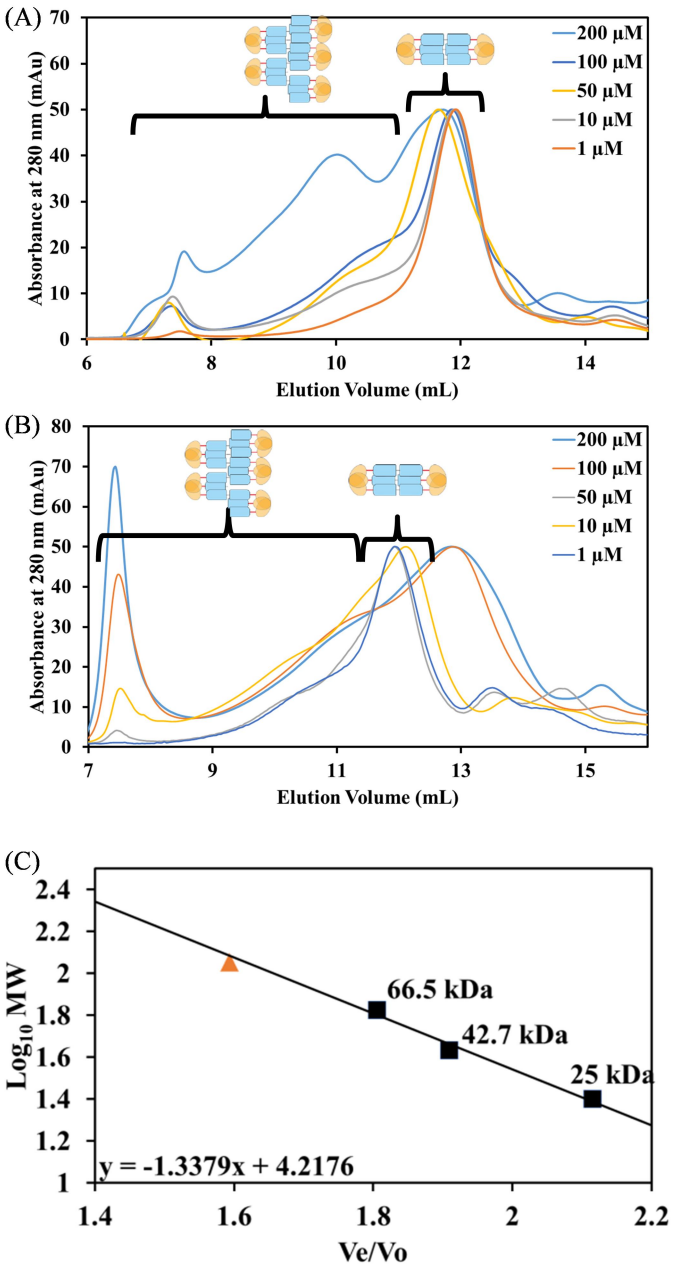


Figure 3.26 Analytical SEC profiles of the purified completely ligated products **(A)** Imp split-intein mediated NCL and **(B)** Gp split-intein mediated NCL (Data collected by Dr. J Wright), at a range of 200 – 1 µM reaction concentrations. 100 µL of concentrated flow-through was injected onto a Superdex 200 10/30 analytical column using standard ligation buffer. **(B)** The product purified from 200 µM and 100 µM ligation reactions were analysed before a filter change on the Superdex 200 10/30. Note: the peak difference of 1 mL of the products purified from 200 µM and 100 µM ligation reactions mediated by Gp split-intein is because the samples were analysed before a filter change on the Superdex 200 10/30. **(C)** The cages Ve/Vo (orange triangle) plotted against their Log_{10} molecular weights on a standard curve. Black squares represent protein standards.

3.5.6 Structure analysis of purified cage products

The purified fully ligated cage structures were characterised by matrix-assisted laser desorption/ionization - time-of-flight (MALDI-TOF) mass spectrometry (MS), circular dichroism (CD) and SEC-small angle x-ray scattering (SEC-SAXS).

3.5.6.1 MALDI-TOF MS

MALDI-TOF MS was performed to confirm the size of the cage product (Section 2.5.5). A strong signal was obtained for the monomeric M4P-CTPR6-M4P product at a mass of 39 kDa (calculated mass of 38 kDa) (Figure 3.27). A peak corresponding to a dimeric $(M4P\text{-}CTPR6\text{-}M4P)_2$ product was also seen at a mass of 78 kDa (calculated mass of 76 kDa). Unfortunately, the complete trimeric cage $(M4P\text{-}CTPR6\text{-}M4P)_3$ product was not observed. This may be due a combination of the limitations of the detector and the inherent method of laser power required for ion formation via MALDI.

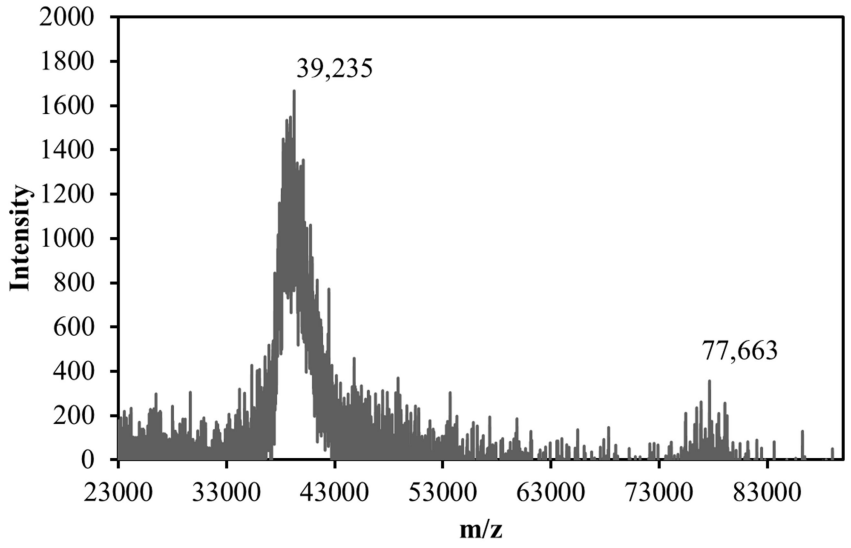


Figure 3.27 MALDI-TOF analysis of cage product. Monomer expected size 38 kDa, observed 39 kDa, percentage difference 3 %. Dimer expected size 76 kDa, observed 78 kDa, percentage difference 2 %. Trimer expected size 113.37 kDa, no further peak was observed.

3.5.6.2 Circular dichroism

Far-UV CD spectroscopy analysis was performed to observe the secondary structure in solution of the purified cage product (Section 0). The purified cage product was compared to a half-cage cap that contains no split-inteins. The resulting CD spectra was calculated and plotted as shown in Figure 3.28. The far-UV CD spectra of the ligated cages show that: (i) they are highly alpha-helical, and, importantly, (ii) have exactly twice the molar ellipticity at 222 nm as that of the half cage caps that do not contain split-intein domains (Figure 3.28A). Figure 3.28B shows the molar ellipticity per residue. The differences in the intensity at 222 nm between the half cage and cage because the number of residues of the cage is not two times of the number of residues of the half cage; the half cage contains a 6-Histidine tag and a FXa cleaving sites, while cage contain the 10 amino acids essential for NCL. The molar ellipticity per CTPR (Figure 3.28C) shows that the extra amino acids present in both half cage and cage did not form more helices. Moreover, the ligation reaction has had no effect on the secondary structure of the CTPR (had not caused any local unfolding).

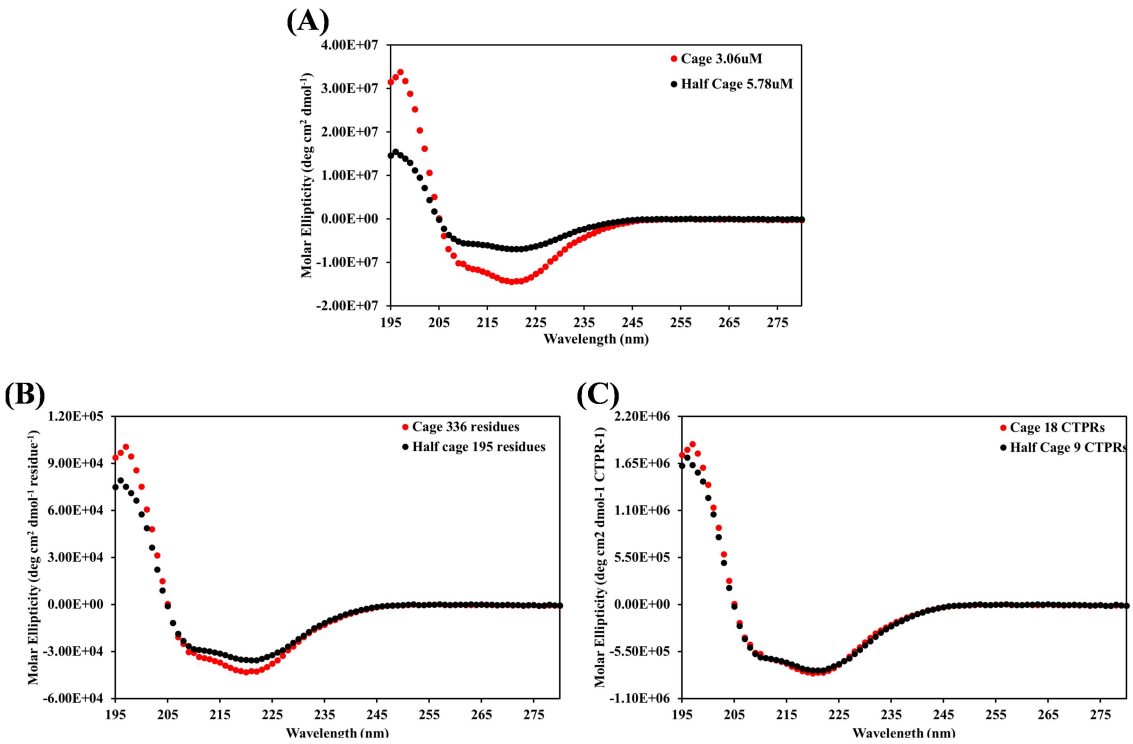


Figure 3.28 Far UV-CD spectra of cage product (red) in comparison to a half-cage cap without split-intein (black) with **(A)** molar ellipticity in deg cm^2 $dmol^{-1}$; **(B)** molar ellipticity in deg cm^2 $dmol^{-1}$ $residue^{-1}$; and **(C)** molar ellipticity in deg cm^2 $dmol^{-1}$ $CTPR^{-1}$.

3.5.6.3 SEC-SAXS

SAXS was carried out to identify the 3D shape of the cage structure. The collected SAXS data was processed and analysed as described in Section 2.5.7.1. Guinier and Kratky plot analysis of the SAXS data confirmed that the purified cages were monodisperse and highly rigid. Moreover, the analysis shows that the cages are non-spherical and elongated with a radius of gyration (Rg) of 3.85 nm and a maximum linear particle diameter (Dmax) of 12.6 nm (Figure 3.29). This is in contrast with the SAXS data of the non-ligated half cage caps, where Kratky plot analysis shows that their structures are highly dynamic. Additionally, the molecular weight of the cages obtained from the SAXS data is in close agreement with that calculated from its amino acid sequence (110.5 kDa versus 113.4 kDa, respectively). Table 3.1 summarises the SAXS parameters of cage products obtained from the analysis.

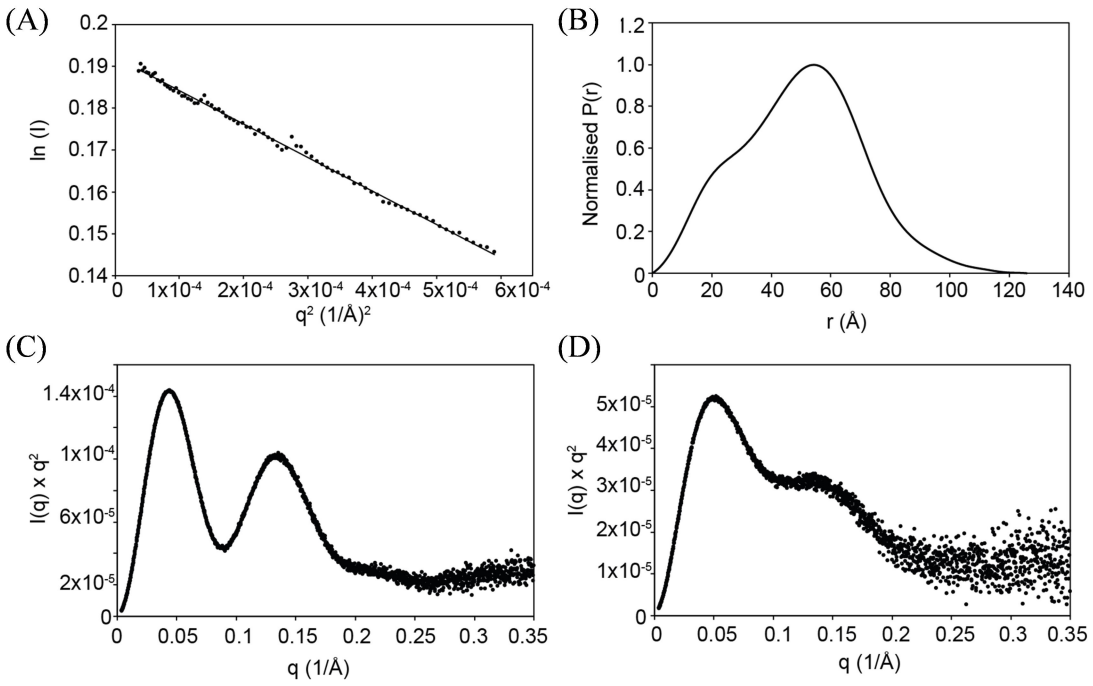


Figure 3.29 Analysis of SAXS of ligated cage and Kratky analysis of half-cage cap. **(A)** Guiner analysis, **(B)** distance distribution functions P(r) and **(C)** Kratky analysis of the SAXS for the ligated cages. **(D)** Kratky analysis for the half-cage cap.

Table 3.1 SAXS parameters obtained from analysis of the purified cage products

SAXS parameters	SAXS cage products
q range ($Å^{-1}$)	0.004 to 0.350
$I(0)$ (Å)	0.134 +/-0.00011
R_g (nm) (from Guinier)	3.85 +/-0.023
R_g (nm) (from $P(r)$)	3.82 +/-0.014
D_{max} (nm) (from $P(r)$)	12.6
Porod Exponent	3.7
MW^{SAXS} (Da)	110,568
$MW^{sequence}$ (Da) of ligated cage	113,058
$MW^{Mass\ Spec}$ (Da) of ligated cage	39,235 (monomeric)

As the SAXS profile of the cage has a number of prominent features it allowed for the determination of its shape to a higher resolution via two different approaches: (i) a SAXS *ab initio* model re-constructed using the program GASBOR (D. I. Svergun, Petoukhov, and Koch 2001) and (ii) a comparison of the experimental SAXS profile to 30 models generated from different possible atomic models of the cage using the program Crysol (D. Svergun, Barberato, and Koch 1995) as described in Section 2.5.7.2.

3.5.6.4 *(i) ab initio* GASBOR modelling

GASBOR reconstructs protein structure by a chain-like ensemble of dummy residues. The *ab initio* GASBOR modelling gave five models that are supported by our biophysical data. Twenty-seven solutions were discarded when, for example, the CTPR/M4P domains would be required to fit protein density envelopes by either adopting non-native conformations or by ligating in a nonsensical formation (discarded examples are shown in Figure 3.30 C-F). The five biophysically relevant solutions were averaged with DAMAVER (Volkov and Svergun 2003) to produce a final model with excellent fit to the data (χ^2 = 1.06) (Figure 3.30 A-B). Excitingly, this final *ab initio* model shows that the ligated cages form our intended structure with protein density that closely resembles the designed trigonal bipyramidal structure and, importantly, a central hollow cavity (M4P oligomer at the "primary" vertices and the CTPRs forming the three cage sides).

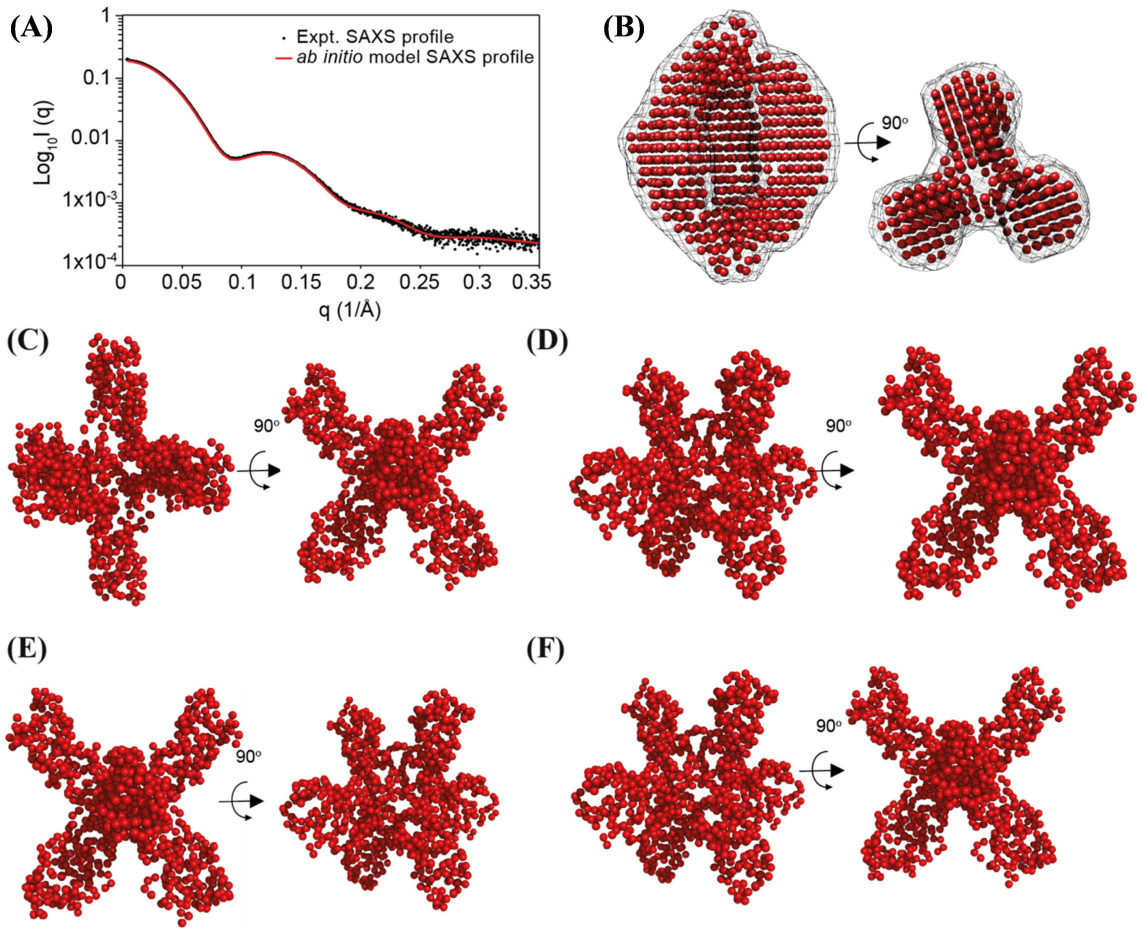


Figure 3.30 *ab initio* model generated by GASBOR. **(A)** Experimental SAXS profile (black dots) of ligated cages overlaid with *ab initio* GASBOR generated model SAXS profile (red line). **(B)** Two orientations of the *ab initio* GASBOR generated model. **(C–F)** Discarded *ab initio* models obtained from experimental SAXS of ligated cage using Gasbor. Solutions were discarded when, for example, the CTPR/M4P domains would be required to either adopt non-native conformations or be ligated in an impossible manner to fit the calculated protein density envelopes.

3.5.6.5 (ii) Comparison of the experimental SAXS profile to models

30 atomic models were generated with slightly different orientation and sizes compared to the SAXS profile using Crysol. Interestingly, all of the generated atomic models that had the same arrangement of protein domains and the central hollow cavity were found to recapitulate the experimental SAXS profile (Table 3.2). Those that have expanded cages or that did not contain a central hollow core gave profiles that were very different from the experimental data. The model that produced the closest fit between experimental and generated SAXS profiles used a continuous CTPR6 as the cage sides (Figure 3.31 A). Here, the CTPR3 modules from the half cages dock upon ligation to form a single CTPR6 superhelix, rather than simply two linked CTPR3 domains like beads on a string. This would account for the increased rigidity of the cage in contrast to the dynamic half cage caps. The final χ^2 value between the model and experimental SAXS profiles was 1.66, with only a small discrepancy at the highest resolution SAXS data (suggesting an ambiguity between the modelled and exact rotation of the CTPR sides and their packing relative to their M4P vertices) (Figure 3.31 B). This model also fits extremely well into the *ab initio* GASBOR model envelope (Figure 3.31 C).

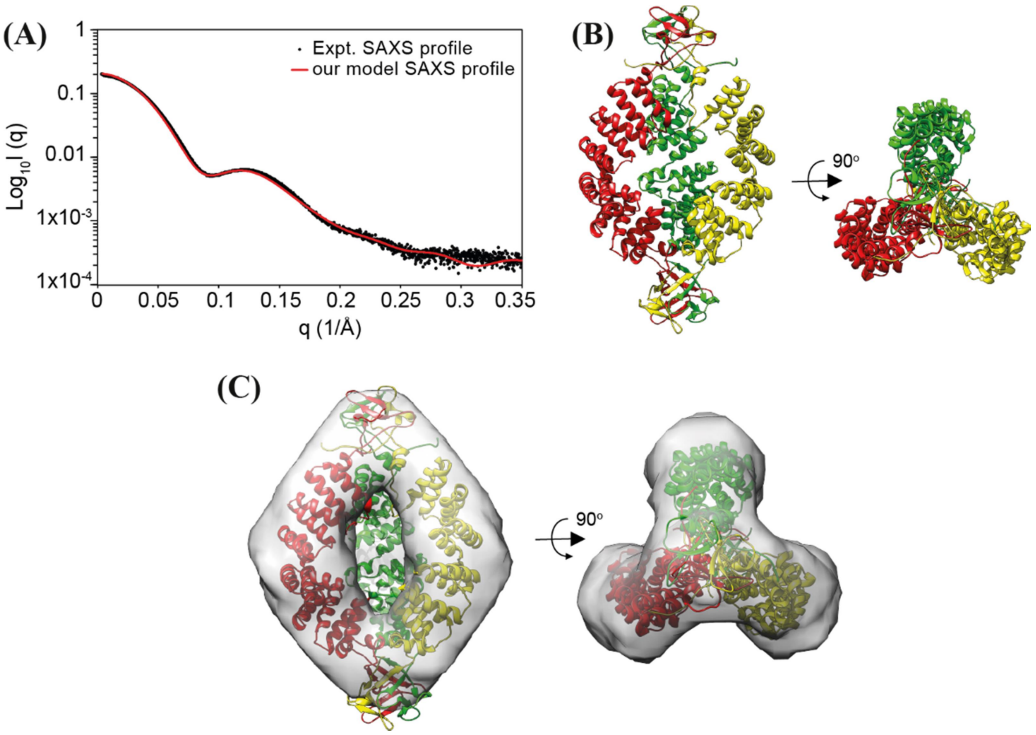
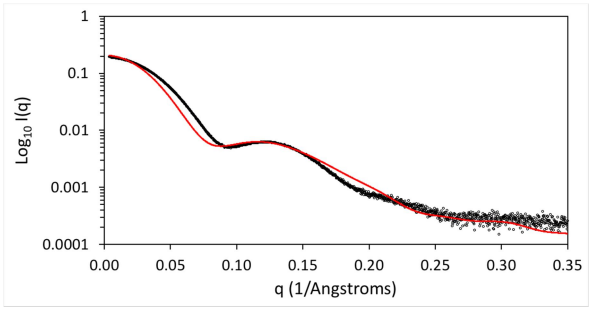
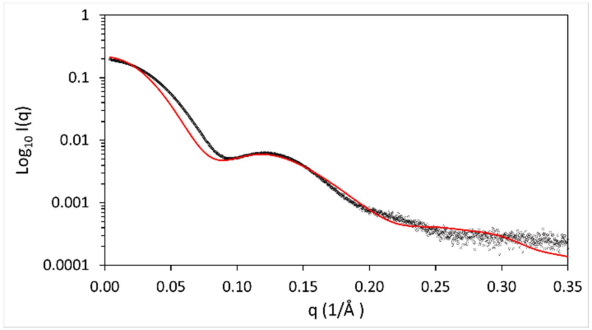
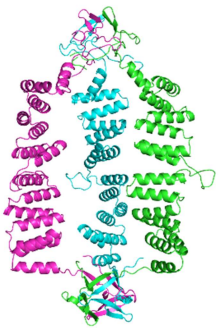
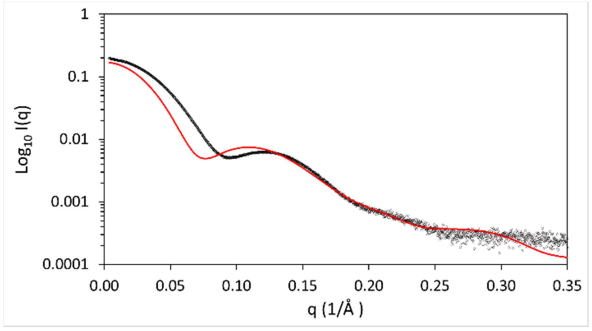
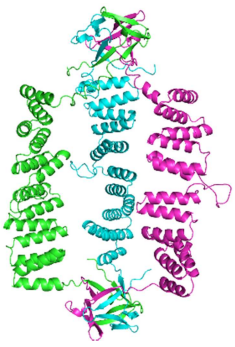
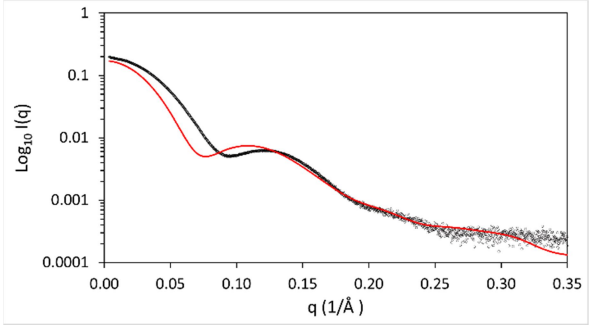
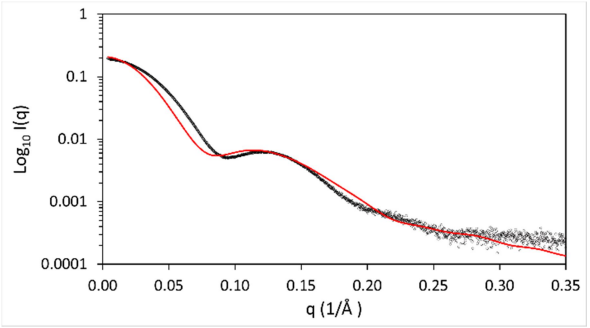
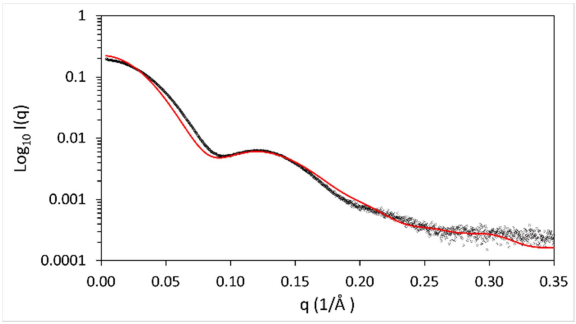
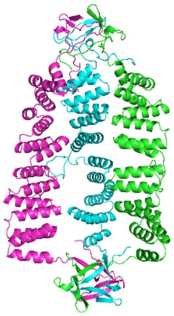
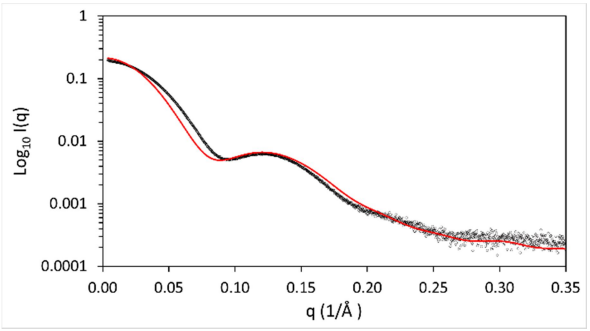
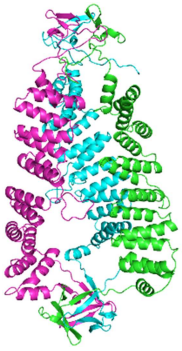
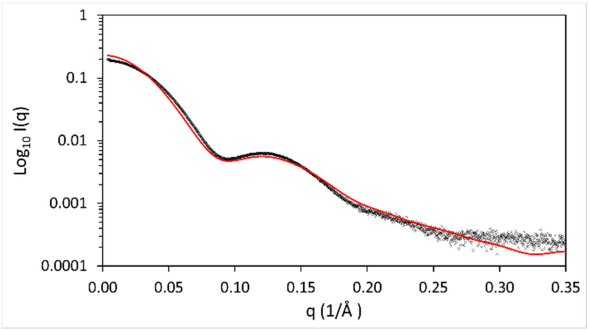
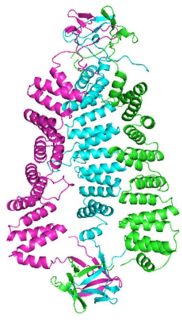
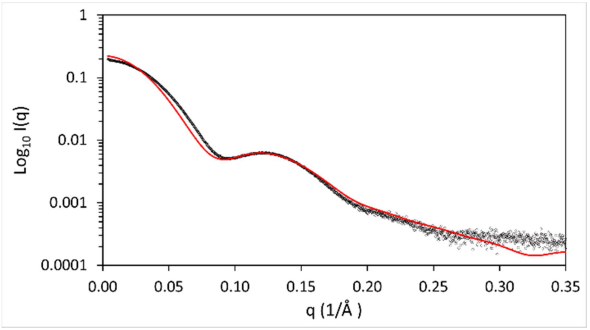
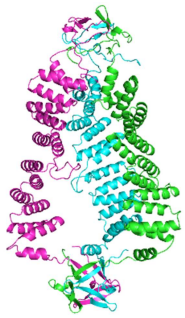
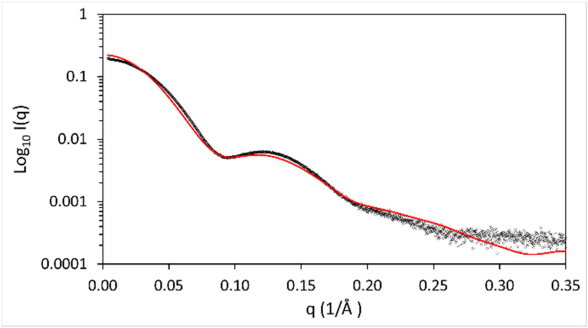
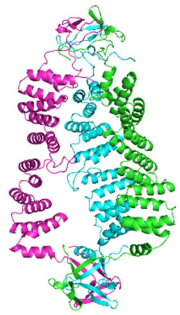
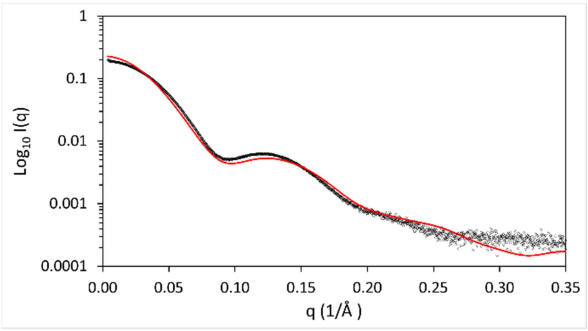
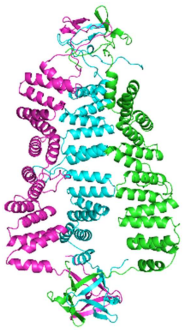
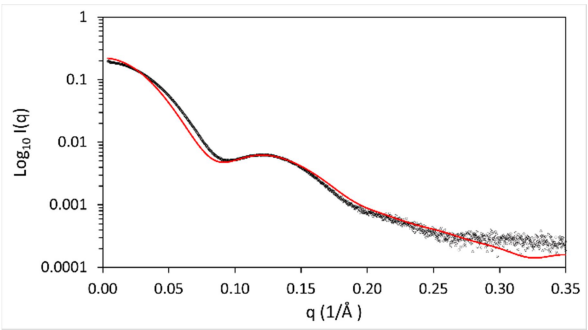
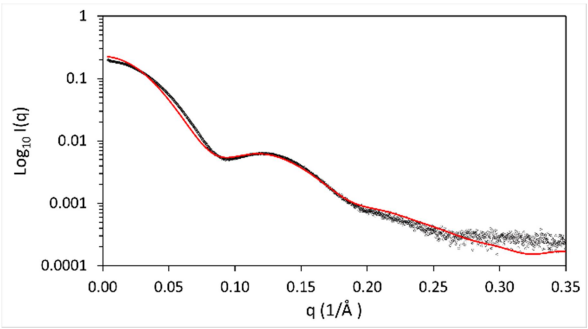
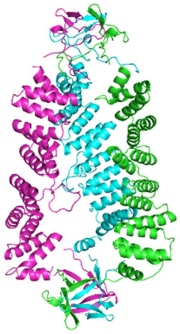
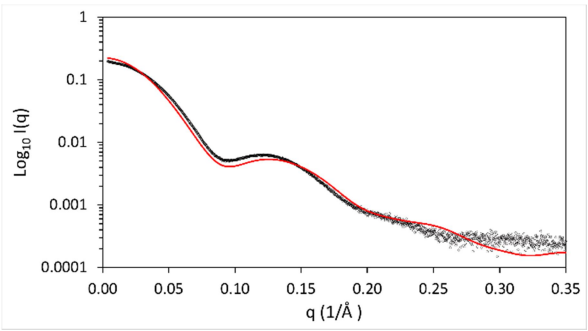


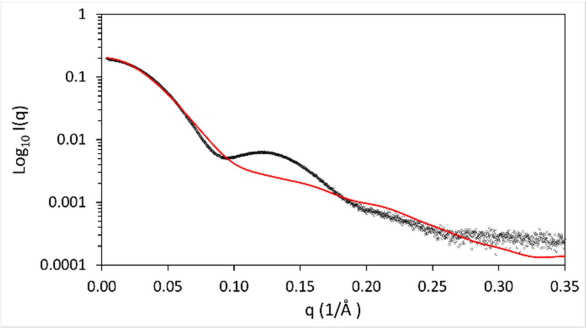
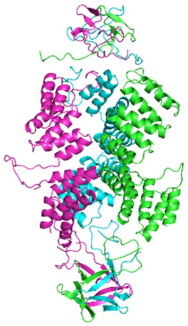
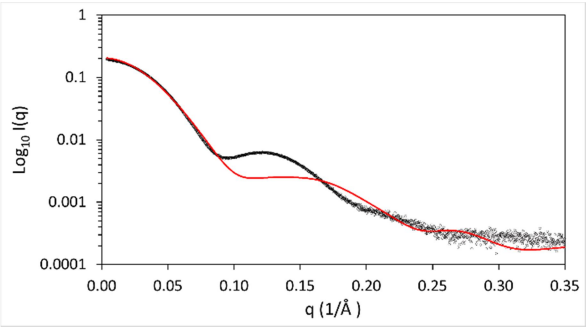
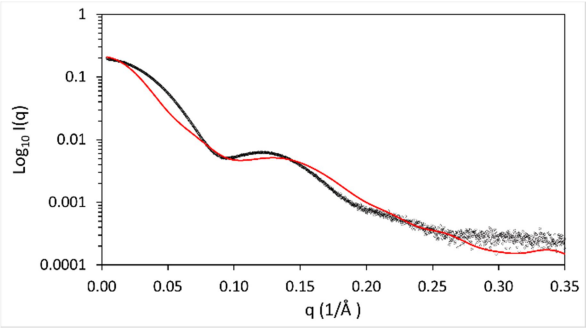
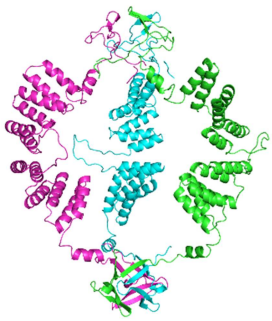
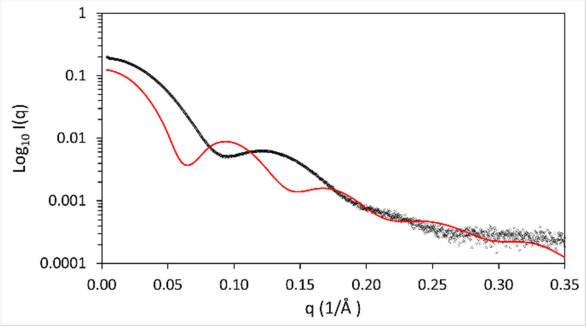
Figure 3.31 Comparison of the 'best-fit' model to the *ab inito* model and the experimental SAXS profile. **(A)** Experimental SAXS profile (black dots) of cage products overlaid with "best-fit" cage atomic model SAXS profile (red line). **(B)** Two orientations of the "best" ligated cage atomic model. **(C)** Two orientations of the "best" ligated cage atomic model superimposed into the *ab initio* GASBOR generated model.

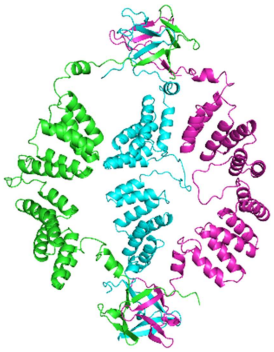
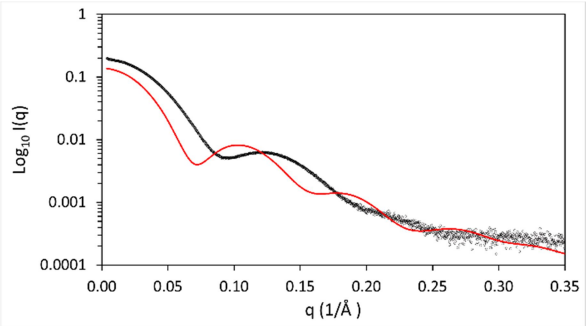
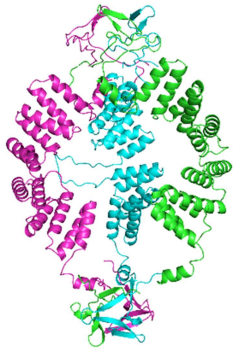
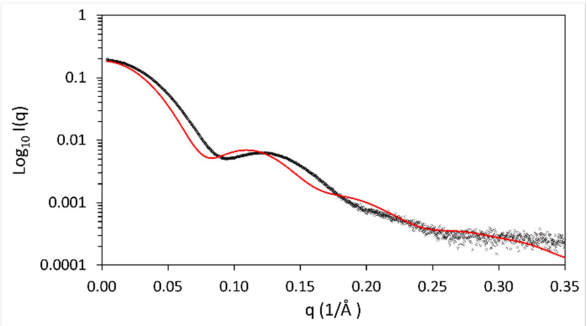
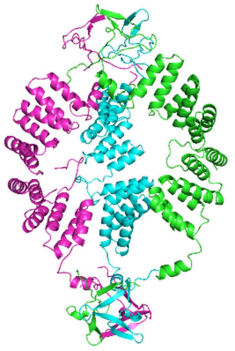
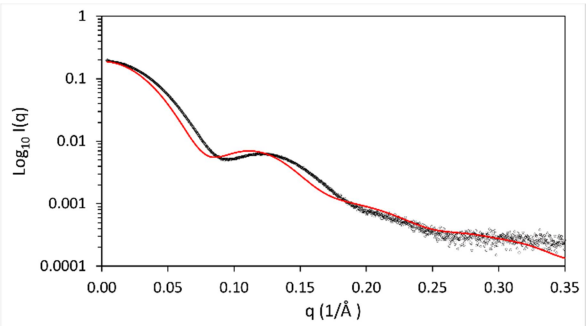
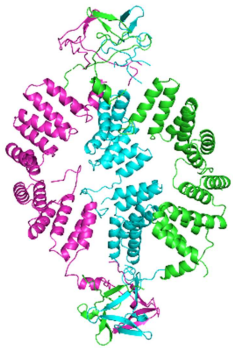
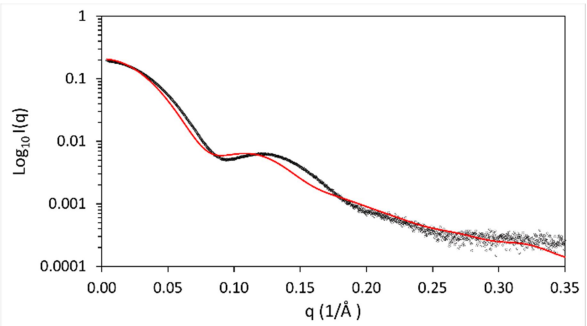
Table 3.2 Comparison of experimental cage SAXS profile and calculated SAXS profile from atomic models of cages.

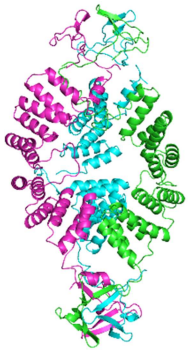
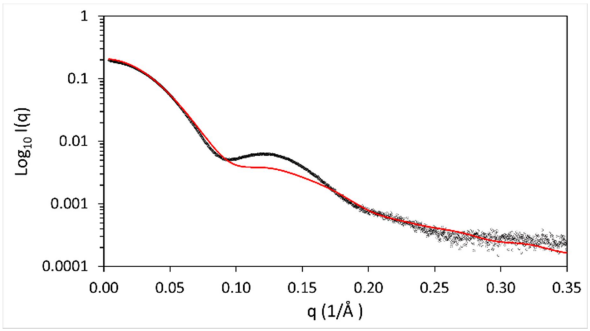
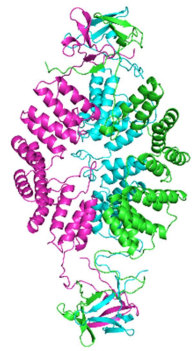
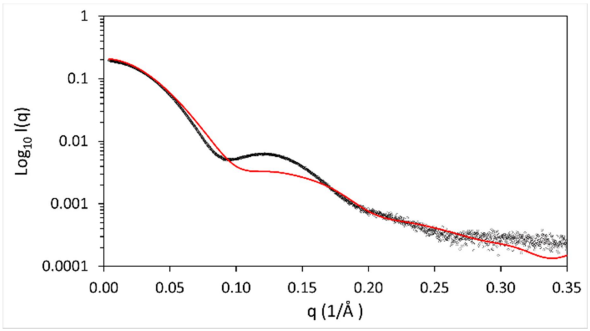
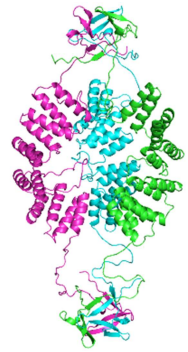
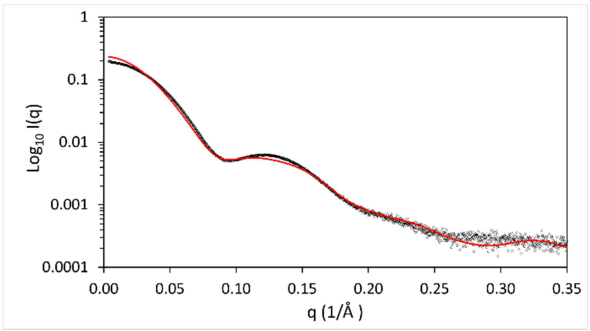
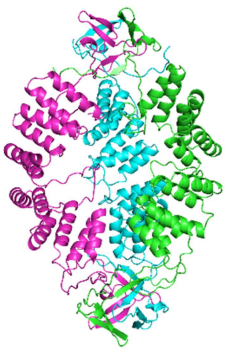
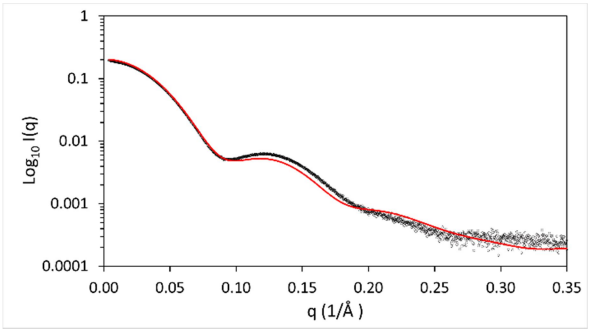
Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Cages with a central cavity but CTPR domain sides not docked		
	5.6	
	5.4	
	8.8	
	7.9	

Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Cages with a central cavity and CTPR domain sides not docked contd.		
	5.6	
	3.5	
Asymmetry Cages, with central cavity but CTPR domains sides not docked		
	3.6	
	3.8	
	4.4	

Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Asymmetry Cages, with central cavity but CTPR domain sides not docked contd.		
	4.7	
	4.6	
	4.7	
	3.9	
	5.0	

Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Cage with "crossed" CTPR domain sides and no cavity		
	9.9	
Cage with docked CTPR domain sides and no cavity		
	8.6	
Expanded cages		
	8.1	
	18.5	

Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Expanded cages, contd.		
	13.9	
	6.4	
	5.0	
	4.1	

Model	χ^2	Comparison of Experimental Ligated Cage SAXS profile (black circles) and model SAXS profile (red line)
Cages where the CTPR sides are docked together into a CTPR6 structure		
	4.3	
	6.7	
	1.8	
	3.2	

3.5.7 Summary

Excitingly, the 3rd generation recombinant proteins of the Imp and Gp split-inteins successfully ligated the half-cage caps in 1 M urea buffer condition with high reaction rates and yields. The various ligated products and excised split-inteins can be easily separated in a two-step process to give homogeneous ligated discrete cages. The yield of discrete cages can be influenced by the concentration of the reactants and the split-intein, which mediates the assembly. For example, it is obvious that lowering the concentrations of reactants and using the Imp-mediated ligation produce more completely ligated discrete cages.

Importantly, when the discrete cages were characterised by MS, CD and SAXS it showed that our design produced the expected trigonal bipyramidal cages with a central hollow cavity. Interestingly, the closest fit model shows that the docking interface between the CTPR3 domains is able to form a continuous superhelix (similarly to CTPR6) despite the 10 amino acids linker between them. Table 3.3 summarises the SAXS profile of the *ab initio* model and 'best' atomic model to the experimental data. Both experimental data and atomic modelling concluded that the ligated cages form an open shell, with a central hollow cavity, that closely resembles the intended designed trigonal bipyramidal structure. Both models showed a closed cage with apertures of ≈ 35 Å between each side at the widest point and encloses a central cavity of ≈ 70 Å by 55-60 Å.

Table 3.3 Comparison of models to experimental SAXS profile

Comparison of Model to Experimental Determined SAXS Profile	
ab initio Gasbor model χ^2 fit	1.06
ab initio Gasbor model NSD	1.02
"Best" Atomic model χ^2 fit	1.66

3.6 Conclusion

In this chapter it has been shown that genetically programmed NCL can be successfully utilised to assemble modular proteins designed with simple geometric symmetry into user-defined protein cages. Through an iterative design process a high yielding system was developed that used split-intein domains, coupled with a two-step purification strategy, to produce homogenous samples of discrete protein cages. To investigate the limits of the system, Imp and Gp split-inteins were trialled separately to drive two-component assembly of the half-cage modules (trefoil vertex and CTPR sides). Significantly, this process generated the expected 113 kDa square bipyramidal structure.

The combined properties of the split-inteins fusion and reaction/purifications yields show that the Imp mediated two-component system could be easily expanded to co-expression and assembly *in vivo*. The use of extendable CTPR sides additionally provides a method for loading cargo into the central hollow cavity of the nanostructures by using a binding module as the linker (Chapter 5). In conclusion, this assembly system provides a more general route to producing protein cages that avoids many time-consuming and system-specific processes (for example, those requiring computational design). No bioconjugation, chemical modification or post-ligation refolding steps were required, and only a short sequence was inserted at the point of the NCL.

4 Design of Controlled Fibre Assembly

4.1 Introduction

This chapter investigates the use of orthogonal split-inteins (Chapter 3) as a driving force to create a stepwise protein fibre extension system. Two different methodologies were employed: (i) tethered linker synthesis - whereby the growing product is eluted from immobilised fusion proteins via affinity tags and (ii) in solution synthesis - whereby protein fusions are reacted in solution and the product of each step is affinity purified. Finally, the limits of both syntheses were delineated (*i.e*. the number of extensions possible).

4.1.1 System design

As with Chapter 3, the system is based on recombinantly expressing a number of fusion proteins. These were termed “caps” and “linkers”. Caps contain the protein to be assembled (POI) expressed as a fusion with a single split-intein half either at its N or C terminus. In contrast the linkers contained the POI sandwiched between two orthogonal split-intein halves (Figure 4.1A). Cloning, expressing and purifying different proteins in these constructs produced initially inert fusions. Mixing with a compatible fusion produced irreversible stepwise assembly, driven by NCL, and enabled directional construction of poly-proteins with specific compositions and spatial arrangements. The fabrication process began with a cap fusion reacting with a linker and allowed the build-up of the protein nanostructures either from the N or C-terminus. Figure 4.1B, showed an exemplar construction from the N-terminus.

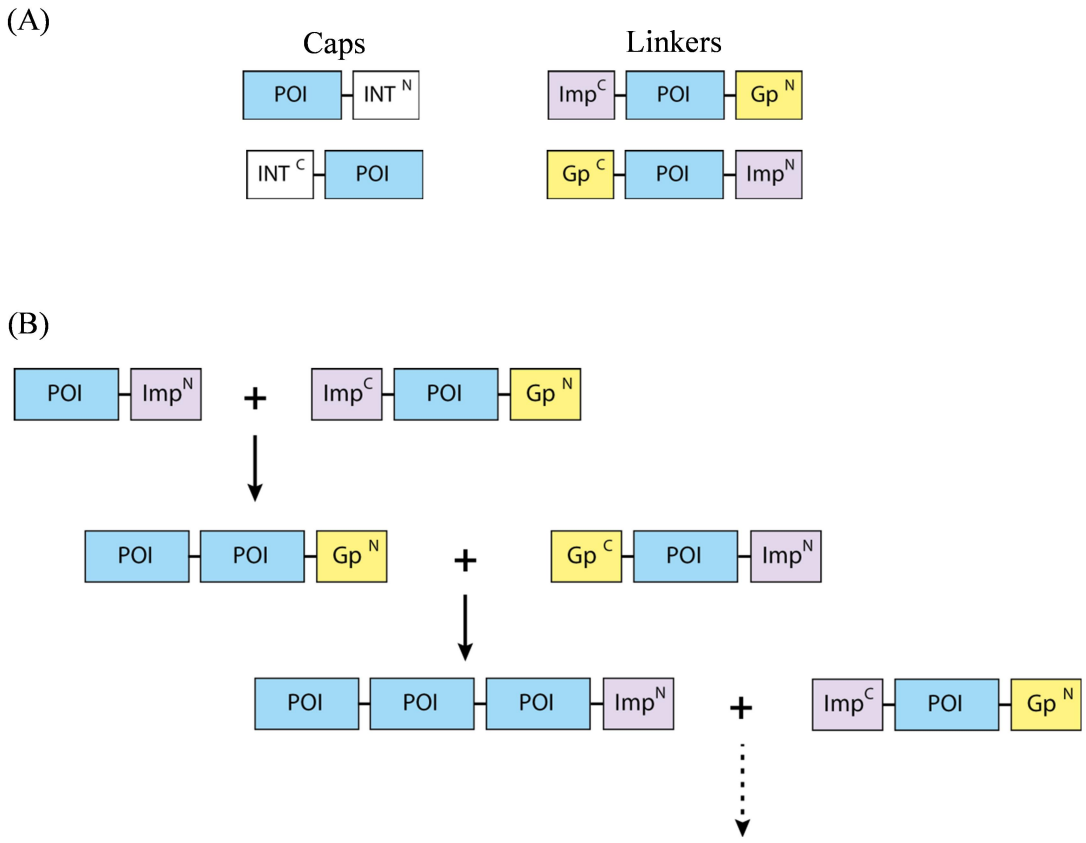


Figure 4.1 **(A)** The recombinant fusions required for stepwise extension: Caps and Linkers, where INT can be either Imp or Gp split-intein. **(B)** Schematic diagram of fibre extension. As an example, sequential linker extension from CTPR3-Imp^{N} is shown. However, extension could be produced with any cap.

4.2 Previous work and motivation

4.2.1 Mxe Gyr A intein mediated stepwise extension

The Main laboratory has shown that a similar system, using the Mxe GyrA (MxGA) intein as an assembly driver, produces head-to-tail orientated fibres in both one-pot and sequential stepwise fabrications (Jonathan J. Phillips, Millership, and Main 2012; Harvey, Itzhaki, and Main 2018). The stepwise assembly process allowed the controlled extension of protein modules with individual reactions producing a yield of 80 % after 16 hrs. However, fibre extension can only be initiated from the N to C-termini and the purification of the ligated product of each step caused a loss of up to 50 %. Therefore, the combined yield after each stepwise addition is ~50 %, suggesting that 3 to 4 sequential ligations joining 4 to 5 protein modules together is the realistic limit of the system.

4.2.2 Split-inteins mediated stepwise extension

Given the limitations of the MxGA intein system, Dr. J. Wright from the Main laboratory redesigned the system as described in Section 4.1.1 to use the orthogonal split-inteins described in Chapter 3 *i.e.* Imp and Gp. These react faster, are completely orthogonal and do not require activation via protease / reducing agent mediated cleavage. He used the Main laboratories model protein of choice, CTPR3 (a protein consisting of 3 consensus tetratricopeptide repeats), as the protein to be assembled and constructed a minimal system of one cap and two linkers. He then explored both product / linkers tethered stepwise extension and stepwise solution extension.

4.2.2.1 Product tethered stepwise extension:

To enable product tethering, the initial cap fusion was modified to include a Twin-StrepII tag (TStr) at the N-terminus. When mixed with Strep-tactin resin, the affinity tag binds, thus immobilising it. Then, compatible linkers can be added sequentially and left to react with the bound cap to produce bound products (Figure 4.2A). Figure 4.2B shows a SDS-PAGE of the products after each round of a three-step extension.

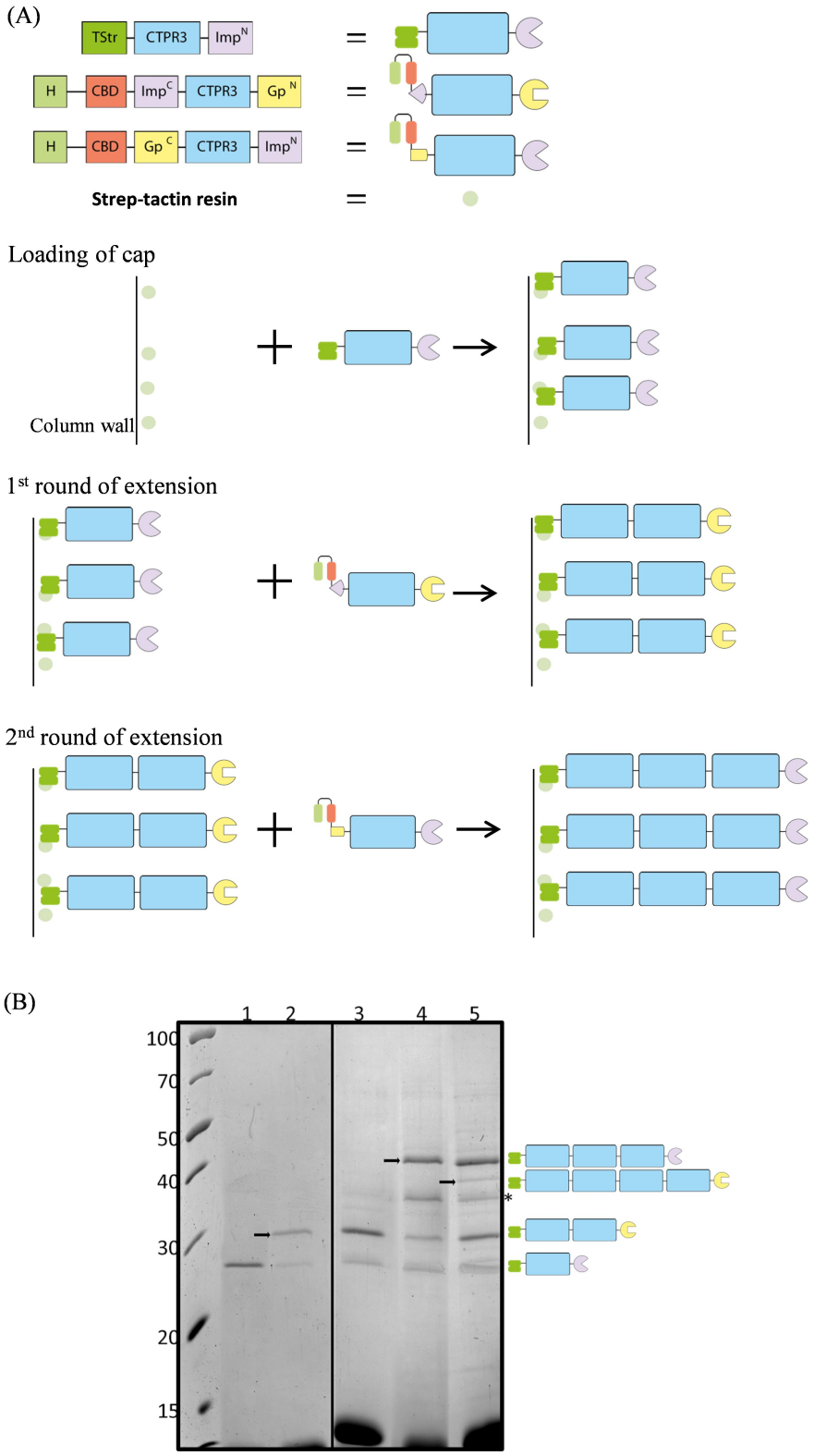


Figure 4.2 **(A)** Schematic of iterative Twin-StrepII tag tethered product reactions. **(B)** Following three rounds of alternate linker additions, H-CBD-Imp^{C}-CTPR3-Gp^{N} and H-Gp^{C}-CTPR-Imp^{N}-CBD to the anchoring domain TStr-CTPR3-Imp^{N}. All samples are resin samples from Strep-tactin resin. Lane 1, Post loading TStr-CTPR3-Imp^{N}; Lanes 2 and 3, first round linker addition; Lane 4, second round linker addition; Lane 5, third round linker addition. Black arrows represent each round's reaction product '*' represents unreacted linker components. Single green rectangle represents 6-Histidine tag; double green rectangle represents twin-strep tags; green circle represents strep resins; blue rectangle represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein. (Data obtained by Dr. J. Wright)

From analysis of the SDS-PAGE, it is clear that the stepwise extensions were successful. However, heterogeneity of the product increased after each round of extension. This is because the reaction yields were not 100 % and thus, the unreacted cap/product can later be extended when further compatible linkers were added (*i.e*. unreacted cap from the first round was extended in the third round).

4.2.2.2 Linker tethered stepwise extension:

Next, linker tethered synthesis was trialled. Here, instead of a cap fusion being immobilised, the linkers were (Figure 4.3). The linkers were immobilised via their chitin binding domains (CBD) bound to chitin resin (H-CBD-Imp^{C}-CTPR3-Gp^{N} and H-CBD-Gp^{C}-CTPR3-Imp^{N}). In each case, when a stepwise reaction takes place the excised split-inteins remained bound to the resin and the product is eluted. By switching between resin loaded with linker fusions containing different split-inteins, the fibre can be extended multiple times (Figure 4.3A & Figure 4.3B).

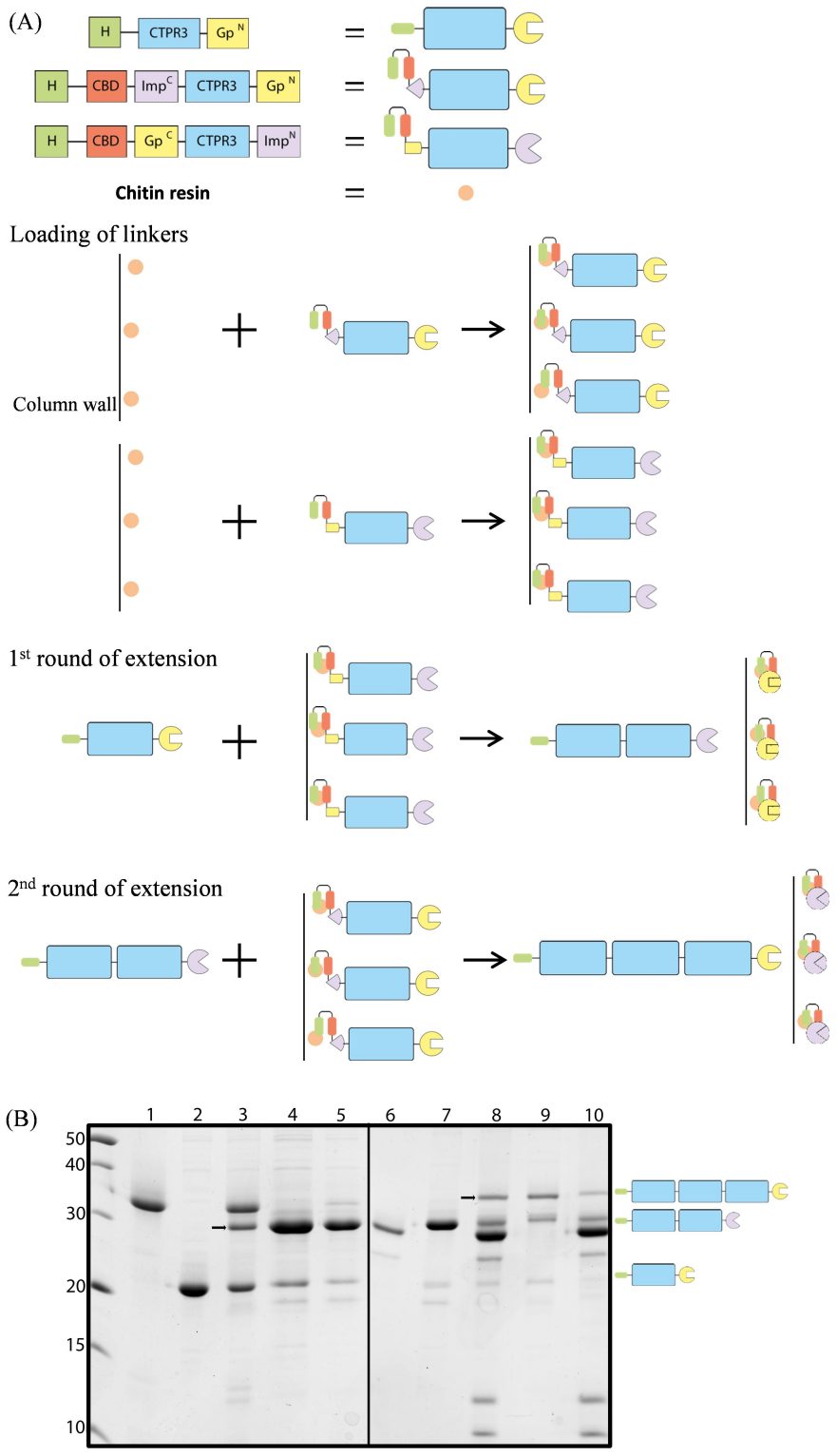


Figure 4.3 **(A)** Schematic diagram of the tethered linker product extension. **(B)** Composite of two gels for clarity, aligned at the border. Lane 1, resin sample of loaded H-CBD-Gp^{C}-CTPR3-Imp^{N}; Lane 2, cap CTPR3-Gp^{N}; Lane 3, resin sample of first linker and cap - 30 mins timepoint; Lane 4, elution from first linker resin 1 hr post reaction; Lane 5, wash from chitin resin; Lane 6, resin sample of loaded H-CBD-Imp^{C}-CTPR3-Gp^{N}; Lane 7, pooled first reaction wash and elution product; Lane 8, resin sample of second linker and first product reactants - 1 hr timepoint; Lane 9, elution from second reactant resin 2 hr post reaction; Lane 10, second reactant resin sample post elution and wash. Black arrows represent the product generated after each round. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; orange represents chitin resins; and yellow represents Gp split-intein. (Data obtained by Dr. J. Wright).

Although the stepwise extension was successful, the reaction yield mediated via the Imp split-intein was much lower than via Gp split-intein (60 % and 90 % in 2 hrs, respectively). Moreover, after the second extension, product was observed in the washed resin samples. This suggests that either the Imp split-intein mediated reaction or ligations that react larger reactants may cause precipitation.

4.2.2.3 Stepwise solution extension:

With regards to the stepwise extension in solution, Dr. J. Wright trialled SEC as a method of separation between products and reactants. He performed two rounds of extension without purification (Figure 4.4A & B). Then, SEC was trialled to separate product from the mixture. Moreover, to determine if the ligated product could be functionalised, the CTPR3 in one linker was replaced with a modified CTPR3, the CTPR390 (H-CBD-Imp^{C}-CTPR390-Gp^{N}) - Figure 4.4A. CTPR390 is a three-repeat CTPR that contains a binding pocket which can then bind to a specific pentapeptide sequence. The first step used excess cap (2:1) and the second equimolar linker. The reactions were left to react for 3 hrs. Figure 4.4B is the SDS-PAGE of the reaction. The extension was successful, however the attempt of purifying the final product by SEC was less successful (Figure 1.4C). This was likely due to the elongated nature of the CTPR proteins, causing differences in molecular weight to not be equally reflected in resolution. The low quality of separation by SEC caused ~50 % loss of the product.

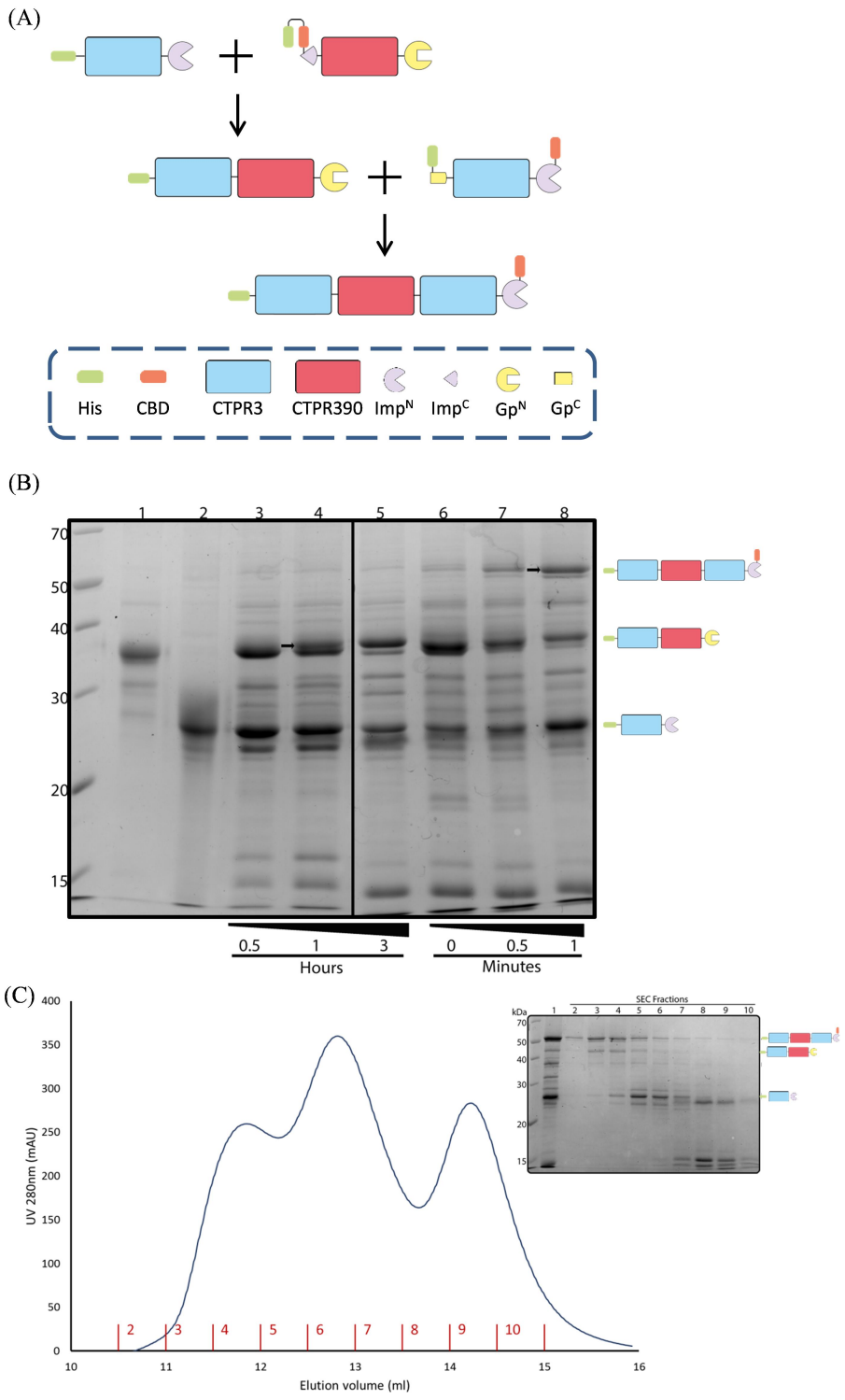


Figure 4.4 **(A)** Schematic diagram representing the three-step solution extension. **(B)** Lane 1, H-CBD-Imp^{C}-CTPR390-Gp^{N}; Lane 2, CTPR3-Imp^{N}; Lane 3, 30 min timepoint sample of the reaction of H-CBD-Imp^{C}-CTPR390-Gp^{N} and CTPR3-Imp^{N} (first reaction step); Lanes 4 & 5, 1 and 3 hr timepoint samples of the first reaction step, respectively; Lane 6, timepoint 0 after the addition of the second linker (H-Gp^{C}-CTPR-Imp^{N}-CBD); Lane 7, timepoint 30 secs of the 2nd reaction step; Lane 8, timepoint 1 min of the 2nd reaction step. Black arrows indicate the generation of the 1st and 2nd reaction products. **(C)** SEC and SDS-PAGE gel of the SEC of the H-CTPR3-CTPR390-CTPR3-Imp^{N} reaction mixture. SDS-PAGE gel: Lane 1, sample of concentrated pre-loaded reaction mixture; Lanes 2-10, SEC fractions corresponding to the graph. (**B** and **C** - Data obtained by Dr. J. Wright). Green represents 6-Histidine tag; blue rectangle represents CTPR3; pink represents CTPR390; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.3 Optimisation of stepwise fibre formation in both linker-tethered and solution synthesis

As can be seen from Dr. Wright's work, optimisation of the stepwise fibre formation via either the linker-tethered or solution synthesis should enable larger assemblies to be constructed. This is because the product after each round of synthesis can be purified to homogeneity. In contrast, product-tethered was investigated, however, optimisation could not increase the yield sufficiently to counteract the increase heterogeneity with each stepwise addition (data not shown). Therefore, linker-tethered and solution synthesis reactions were optimised as follows:

(i) Linker-tethered Synthesis: Firstly, the length of time required for Imp split-intein mediated ligation was increased to 16 hrs. Secondly, instead of initiating the extension with Gp mediated ligation (Gp cap fusion and linker), the extension was initiated with Imp (Imp cap and linker).

(ii) Solution Synthesis: Here, optimum conditions were initially determined by only extending with "Spacer" CTPR domains. The "Binder" CTPR domain that Dr. Wright used required higher denaturant concentration to remain soluble. However, this reduced the overall yield. The system was optimised to enable convergent extension from two caps (Figure 4.5). Finally, the convergent extension was redesigned to enable easier purification of each stepwise extension product.

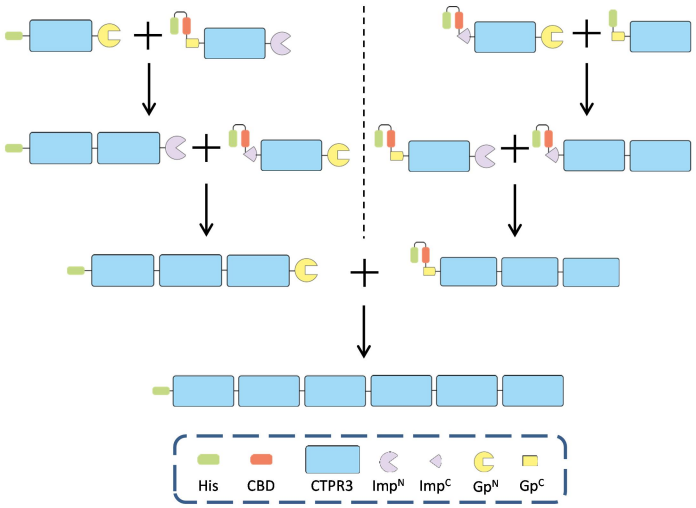


Figure 4.5 Schematic diagram of the making of CTPR18. Two rounds of extension from both the N and C-terminal caps to obtain CTPR9. Then, the two CTPR9s with compatible split-inteins were reacted to obtain CTPR18. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.4 Recombinant expression and purification of protein fusions

To enable the experiments outlined above, the following cap and linker protein fusions were produced and purified as follows:

4.4.1 Recombinant expression and purification of caps: ^{1}H-CTPR3-Imp^{N}, ^{1}H-CTPR3-Gp^{N}, ^{1}H-Gp^{C}-CTPR3 and CTPR3-Gp^{N}-H

All caps except H-Gp^{C}-CTPR3 were successfully expressed and purified natively as described in Section2.3.3.1, with 15 - 40 mg/mL yield and high purity. Whereas, the purification of H-Gp^{C}-CTPR3 required the denatured protocol (Section 2.3.3.2). The denatured recombinant protein was refolded on the column. The protein was eluted in buffer 1 M urea, 50 mM Tris pH 8, 300 mM NaCl and 250 mM imidazole. 5 mM DTT was added into the elution to prevent disulphide formation.

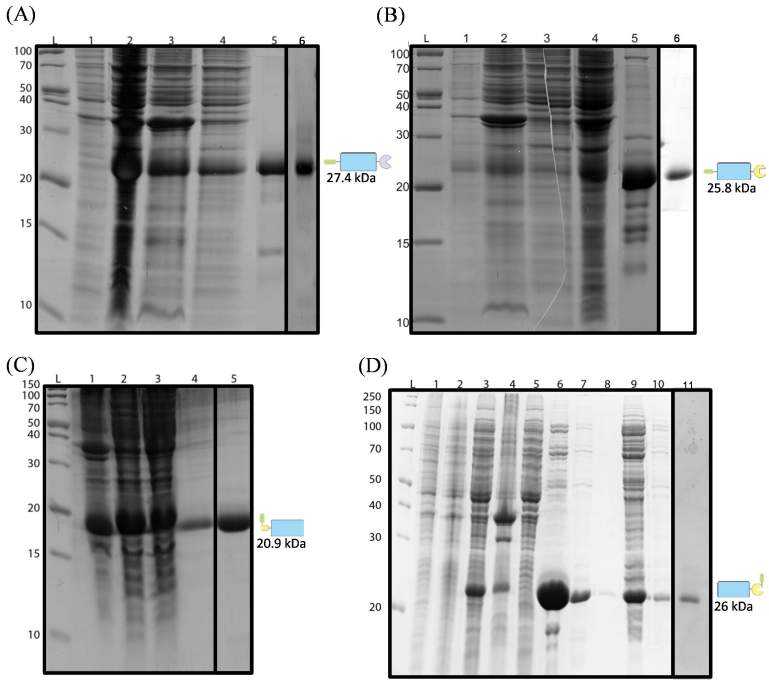


Figure 4.6 SDS-PAGE of the purification of **(A)** ^{1}H-CTPR3-Imp^{N}, **(B)** ^{1}H-CTPR3-Gp^{N}, **(C)** ^{1}H-Gp^{C}-CTPR3 and **(D)** CTPR3-Gp^{N}-H. **(A-D)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. **(A-C)** Lane 1, post-induction culture sample; Lane 2, native insoluble cell lysate; Lane 3, native soluble cell lysate; **(A-B)** Lane 4, flow-through fraction; Lane 5, elution fraction; and Lane 6, purified H-CTPR3-Imp^{N} and H-CTPR3-Gp^{N} respectively; **(C)** Lane 4, elution fraction; Lane 5, purified H-Gp^{C}-CTPR3. **(D)** Lane 1, pre-induction culture sample; Lane 2, post-induction culture sample; Lane 3, native/denatured insoluble cell lysate; Lane 4, native/denatured soluble cell lysate; Lane 5, flow-through fraction; Lane 6-10, elution fractions; Lane 11, purified CTPR3-Gp^{N}-H. (^{1} data obtained by Dr. J Wright). Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

[1] The recombinant expression and purification of H-CTPR3-Imp^{N}, H-CTPR3-Gp^{N}, H-Gp^{C}-CTPR3 were performed by Dr. J. Wright.

4.4.2 Recombinant expression and purification of linkers: H-CBD-Imp^{C}-CTPR3-Gp^{N}, H-CBD-Gp^{C}-CTPR3-Imp^{N}, CBD-Imp^{C}-CTPR3-Gp^{N}-H[1] and H-Gp^{C}-CTPR3-Imp^{N}-CBD

All linkers were successfully purified under denaturing conditions and refolded as described in Section 2.3.3.2. The proteins were eluted in buffer 1 M urea, 50 mM Tris pH 8, 300 mM NaCl, and 250 mM imidazole. 5 mM DTT were added into the elution to prevent disulphide formation. The purification of all proteins yielded 20–55 mg/mL respectively.

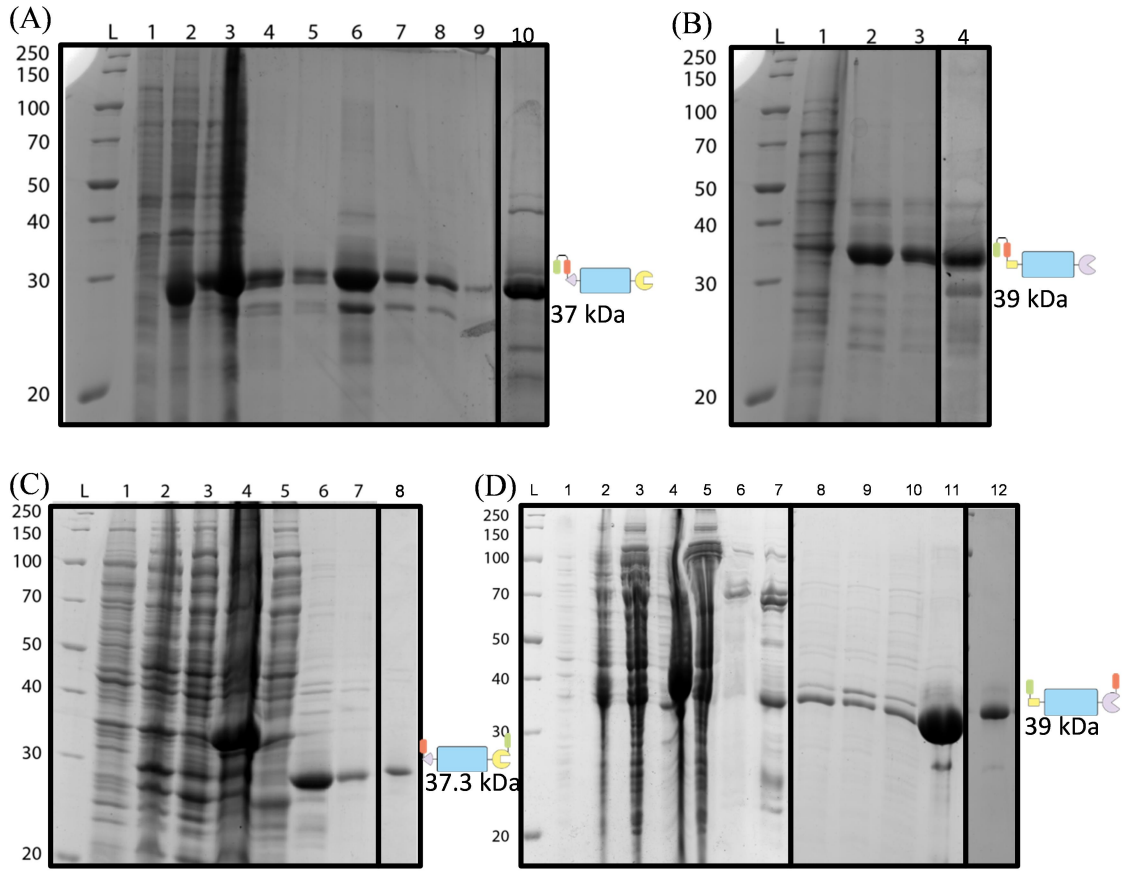


Figure 4.7 SDS-PAGE of the purification of **(A)** H-CBD-Imp^{C}-CTPR3-Gp^{N}, **(B)** H-CBD-Gp^{C}-CTPR3-Imp^{N}, **(C)** CBD-Imp^{C}-CTPR3-Gp^{N}-H and **(D)** H-Gp^{C}-CTPR3-Imp^{N}-CBD. **(A-C)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. **(A)** Lane 1, pre-induction cell sample; Lane 2, post-induction cell sample; Lane 3, flow-through fraction; Lanes 4-9, elution fractions; and Lane 10, purified H-CBD-Imp^{C}-CTPR3-Gp^{N}. **(B)** Lane 1, denatured flow-through; Lanes 2 and 3, elution fractions; and Lane 4, purified H-CBD-Gp^{C}-CTPR3-Imp^{N}. **(C)** Lane 1, pre-induction culture sample; Lane 2, post-induction culture sample; Lane 3, denatured insoluble lysate; Lane 4, denatured soluble lysate; Lane 5, flow-through fraction; Lanes 6-7, elution fractions; and Lane 8, purified CBD-Imp^{C}-CTPR3-Gp^{N}-H. **(D)** Lane 1, pre-induction culture; Lane 2, post-induction culture; Lane 3, native soluble lysate; Lane 4, native insoluble lysate; Lane 5, affinity column flow-through; Lane 6, affinity column wash; Lane 7, affinity column elution fraction; Lane 8, denatured soluble lysate, Lane 9, denatured insoluble lysate; Lane 10, affinity column flow-through; Lane 11, affinity column elution fraction; and Lane 12, purified H-Gp^{C}-CTPR3-Imp^{N}-CBD. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

[1] The recombinant expression and purification of H-CTPR3-Imp^{N}, H-CTPR3-Gp^{N}, H-Gp^{C}-CTPR3 were performed by L. Synchyshyn

4.5 Linker tethered extension optimisation

4.5.1 Reaction Conditions

7 mL of chitin resin was charged with an excess of the required linker fusion. These were thoroughly equilibrated in reaction buffer supplemented with 1 M urea which removed any excess or loosely bound linker (Figure 4.8).

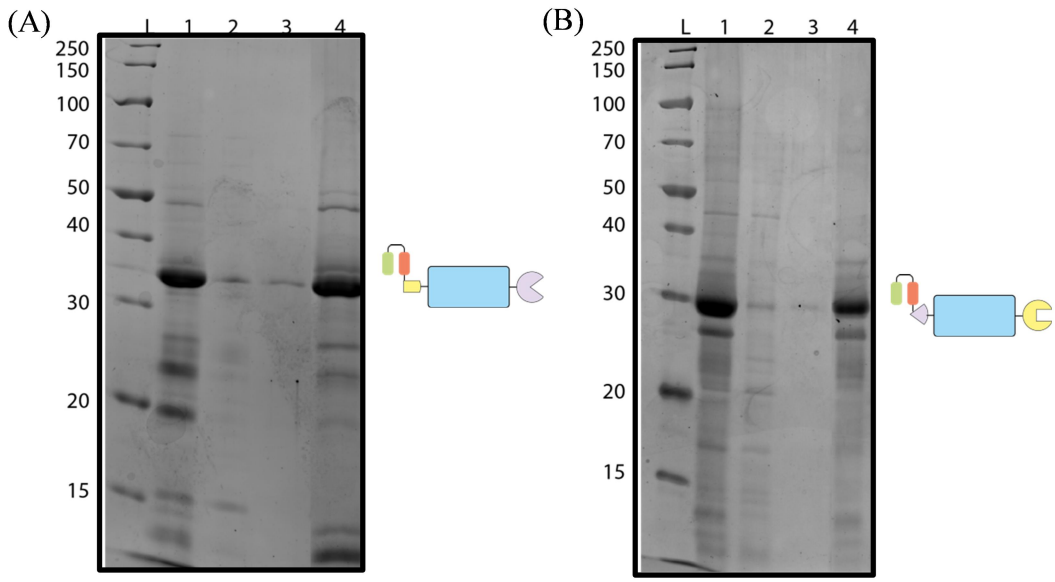


Figure 4.8 The SDS-PAGE of **(A)** H-CBD-Gp^{C}-CTPR3-Imp^{N} and **(B)** H-CBD-Imp^{C}-CTPR3-Gp^{N} bound to a chitin resin column. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, purified linker; Lane 2, flow through of the linker from the column; Lane 3, wash fraction; and Lane 4, resin of the bound protein. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

To initiate the reaction, 5 mL of 50 µM of cap was added to resin charged with a compatible linker. The reaction was incubated with gentle agitation at 25 °C. Once the reaction had reached completion, the flow through of the column was collected and concentrated to 5 mL. This could then be added to resin charged with the next compatible linker to initiate the next stepwise extension. The process could then be repeated to produce iteratively larger assemblies (Figure 4.3A).

4.5.2 Varying Imp split-intein mediated reaction time

For the first round of extension, Dr. J. Wright's protocol was repeated. Here, step one reacted the cap, H-CTPR3-Gp^{N}, with immobilised H-CBD-Gp^{C}-CTPR3-Imp^{N} linker. The reaction was mediated by a Gp split-intein and was left to proceed for 2 hrs at 25 °C. Figure 4.9A shows the SDS-PAGE of the reaction. The reaction yield was calculated using the band intensities of the product and unreacted protein found in the elution (Figure 4.9A-Lane 3) – Section 2.5.2. The reaction yield mediated by Gp split-intein was consistent with Dr. J. Wright's findings, which was 90 % yield.

In order to increase the reaction yield in the second round of extension, the reaction time of the concentrated H-CTPR6-Imp^{N} and immobilised H-CBD-Imp^{C}-CTPR3-Gp^{N} (mediated by Imp split-intein) was extended to 16 hrs. The reaction yield obtained was same as the result obtained by Dr. J. Wright, *i.e.* 60 %.

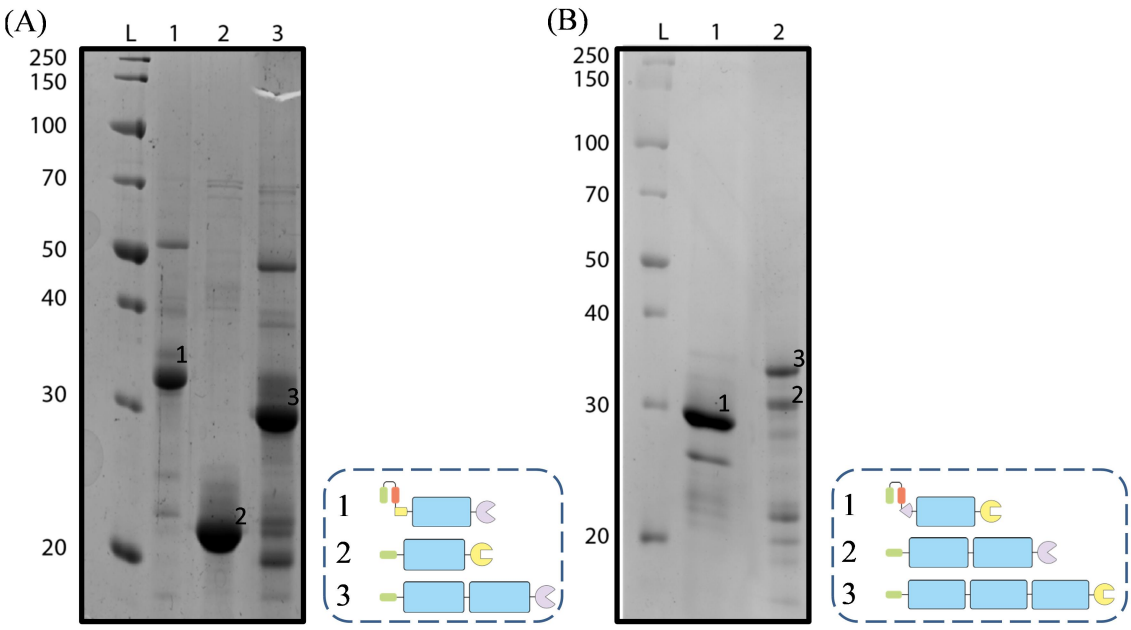


Figure 4.9 SDS-PAGE of the linker tethered extensions. **(A)** First extension, ligation of H-CTPR3-Gp^{N} and immobilised H-CBD-Gp^{C}-CTPR3-Imp^{N}. Lane 1, resin sample of bound H-CBD-Gp^{C}-CTPR3-Imp^{N}, Lane 2, H-CTPR3-Gp^{N}; Lane 3, elution fraction after 1 hr reaction time (expected product, H-CTPR6-Imp^{N}, 38.6 kDa) **(B)** Second extension, ligation of H-CTPR6-Imp^{N} and immobilised H-CBD-Imp^{C}-CTPR3-Gp^{N}. Lane 1, resin sample of bound H-CBD-Imp^{C}-CTPR3-Gp^{N}; Lane 2, elution fraction after 16 hrs reaction time (expected product, H-CTPR9-Gp^{N}, 50.2 kDa). **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.5.3 Stepwise extension initiated by Imp split-intein

Increasing the reaction time of the second step (Imp mediated ligation) failed to increase the reaction yields. This suggests that the Imp split-intein might have lost some reactivity before step 2. Therefore, it was decided to change the sequence of the split-inteins used. Instead of initiating with a Gp-mediated ligation, an Imp split-intein cap (H-CTPR3-Imp^N) was reacted with a compatible linker (H-CBD-Imp^C-CTPR3-Gp^N). In solution this reaction proceeded to 85 % yield within 3 hrs (data not shown). Excitingly when the reaction was performed with the linker tethered to the resin, a similarly high yield of 80 % was obtained (Figure 4.10A). The product was isolated successfully and the 2nd stepwise addition was trialled. This was mediated by Gp split-inteins with resin bound H-CBD-Gp^C-CTPR3-Imp^N linker (incubated for 2 hrs) (Figure 4.10B). This was also successful, with a yield of 80 %. Finally, a third extension was performed with the 2nd reaction product (H-CTPR9-Imp^N) reacted with the initial H-CBD-Imp^C-CTPR3-Gp^N charged resin (Figure 4.10C). A reaction yield of 65 % was obtained but 50 % of the product was lost during purification and concentrating steps. Figure 4.11 shows the product from each stepwise addition.

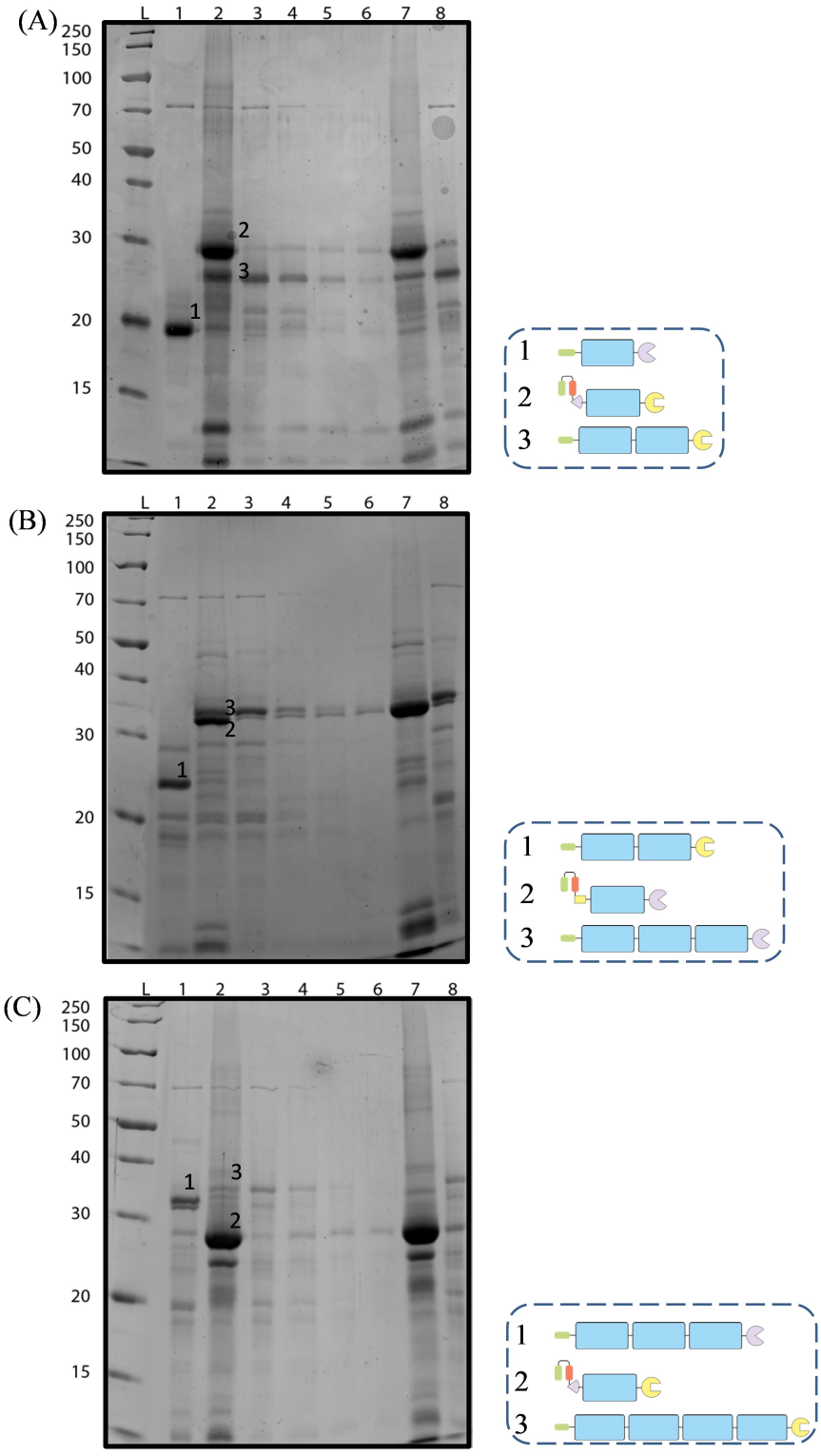


Figure 4.10 The SDS-PAGE of the stepwise extension of the fibres. **(A)** The reaction of H-CTPR3-Imp^{N} and H-CBD-Imp^{C}-CTPR3-Gp^{N}, producing H-CTPR3-CTPR3-Gp^{N} (30 kDa) in 3 hrs. **(B)** The reaction of H-CTPR3-CTPR3-Gp^{N} and H-CBD-Gp^{C}-CTPR3-Imp^{N}, producing H-CTPR3-CTPR3-CTPR3-Imp^{N} (39 kDa) in 2 hrs. **(C)** The reaction of H-CTPR3-CTPR3-CTPR3-Imp^{N} and H-CBD-Imp^{C}-CTPR3-Gp^{N}, producing H-CTPR3-CTPR3-CTPR3-CTPR3-Gp^{N} (54.8 kDa). **(A-C)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, reactant added to the column bound with the linker; Lane 2, resin sample of the bound linker; Lane 3, resin sample after the reaction; Lanes 4-6, the flow-through and wash fraction after the reaction; Lane 7, resin sample of bound protein after wash; and Lane 8, concentrated sample of the flow-through and wash fractions after the reaction. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

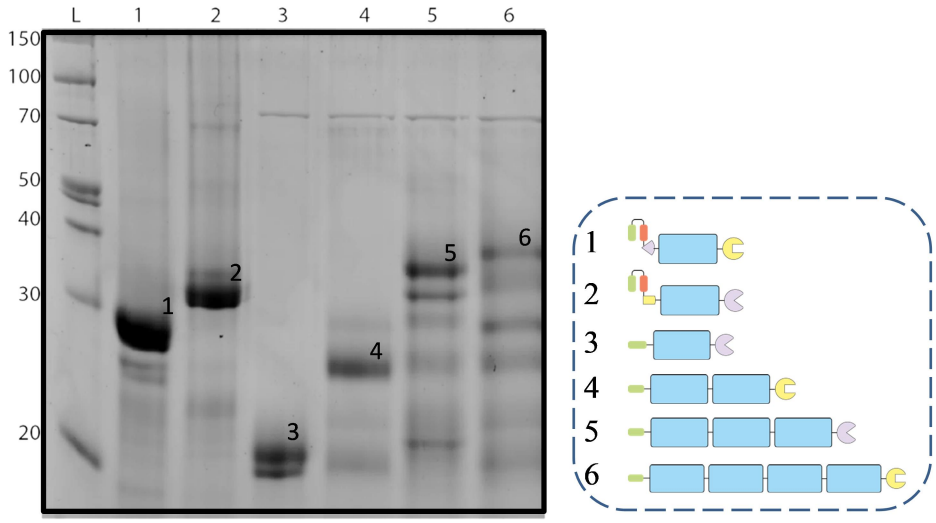


Figure 4.11 The SDS-PAGE of the stepwise extension of CTPR3. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, bound H-CBD-Imp^{C}-CTPR3-Gp^{N}; Lane 2, bound H-CBD-Gp^{C}-CTPR3-Imp^{N}; Lane 3, H-CTPR3-Imp^{N}; Lane 4, 1st round of extended product, H-CTPR3-CTPR3-Gp^{N} (30 kDa); Lane 5, 2nd round of extended product, H-CTPR3-CTPR3-CTPR3-Imp^{N} (39 kDa); and Lane 6, 3rd round of extended product, H-CTPR3-CTPR3-CTPR3-CTPR3-Gp^{N} (54.8 kDa). Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.5.4 Summary – Linker tethered stepwise extensions

Dr. Wright had successfully shown that linker tethered ligations could be successfully used in a 2-reaction sequence to link three proteins together. The first Gp-mediated reaction step produced high yields. However, the 2nd Imp-mediated step only produced 60 % yield. As Imp requires longer than Gp to react, increasing the time for the Imp reaction step was trialled. However, this did not increase the yield higher than the original 60 %. The low yield may be due to Imp split-intein losing its reactivity or aggregating during the first ligation reaction. Hence, the system was redesigned to start the extension with the Imp split-intein. Excitingly, this significantly increased the yield for the Imp step to ~80 %. A second round of extension via Gp split-intein was then successfully achieved with an ~80 % yield. However, when a further Imp mediated round of extension was trialled the yield dropped to 65 %. The reduced in yield, in the third round, occured with the Imp mediated split-intein again. One reason for the decrease in yield could be that the size of the protein becomes either sterically hindered or more likely to aggregate on the resin or losing its reactivity. No matter the exact cause of the reduced in reaction yield in the 3rd step, and the loss of product during the concentrating step, it can be concluded that the tethered linker synthesis is limited to two rounds of extension.

4.6 Stepwise solution extension

From Dr. Wright's work and the previous sections it is obvious that 3 to 4 ligations are the limit of Gp / Imp split-intein mediated assembly. However, we can assemble from both the N and C-termini. Therefore, the effective size of product can be increased by extending from both termini and using the final ligation step to ligate both together. For example, three rounds of extension of CTPR3 from either N or C termini produces a CTPR12 protein. Whereas extending from both N / C termini and then ligating both in the 3rd round could produce a CTPR18 protein (Figure 4.5). Moreover, the increased size differential should enable easy SEC separation of the final product.

4.6.1 1st Generation

Our first convergent solution assembly, Figure 4.5, uses Gp-mediated ligation in step 1, followed by Imp-mediated ligation in step 2. Both products are then combined in step 3 and ligated together via a final Gp-mediated ligation. Steps 1 and 2 were performed using a ratio of 2 : 1 : 2 (cap : 1st linker : 2nd linker), where the concentration of the cap protein was 100 μM. An excess of cap in step 1 and linker in step 2 were used to drive the generation of the greatest possible. All reactions were performed under standard reaction conditions (Section 2.4.2) with a reaction buffer supplemented with 1 M urea. The reaction time for Gp and Imp mediated ligation were 1.5 hrs and 3 hrs, respectively. Figure 4.12 shows the results from the 3 rounds of extension.

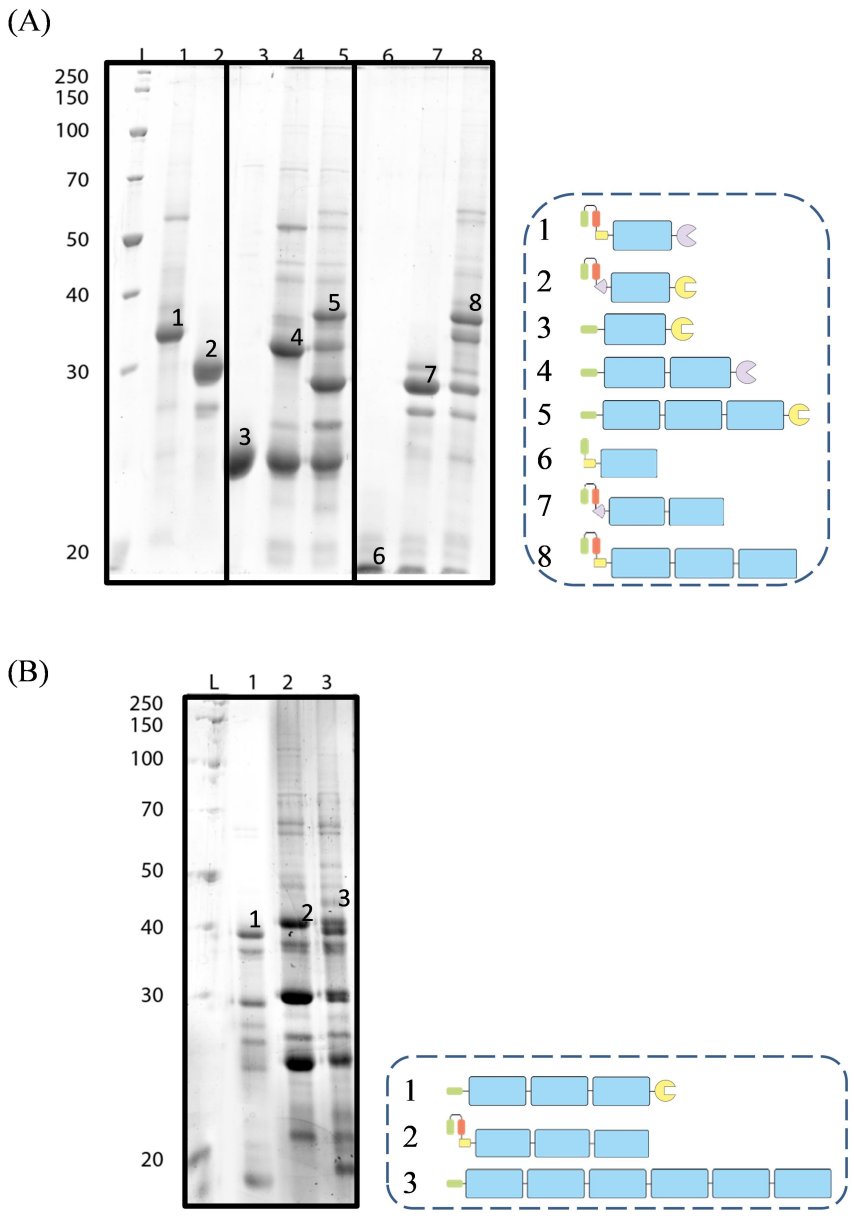


Figure 4.12 SDS-PAGE of the extension of fibres in solution. **(A)** Extension from both N and C-terminal caps. Lane 1, H-CBD-Gp^{C}-CTPR3-Imp^{N}, Lane 2, H-CBD-Imp^{C}-CTPR3-Gp^{N}; Lane 3, H-CTPR3-Gp^{N}; Lane 4, H-CTPR6-Imp^{N} (41.7 kDa); Lane 5, H-CTPR9-Gp^{N} (53.4 kDa); Lane 6, H-Gp^{C}-CTPR3, Lane 7, H-CBD- Imp^{C}-CTPR6 (40.6 kDa); and Lane 8, H-CBD-Gp^{C}-CTPR9 (54.2 kDa). **(B)** Ligation of extended caps. Lane 1, H-CTPR9-Gp^{N}; Lane 2, H-CBD-Gp^{C}-CTPR9; Lane 3, the reacted mixture after 1.5 hrs. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.6.1.1 Summary

The performance of Gp and Imp split-inteins were consistent. A high yield was gained from Gp-mediated ligation, whereas Imp-mediated ligation gave a slightly lower yield. The final ligation produced very little CTRP18. This was due to the unwanted ligation of extra reactants. To solve this problem, the purification of the product was required after each round of ligation. Hence, the system was redesigned as per the next section.

4.6.2 2nd Generation

To enable easier purification of each step and to reduce unwanted "side reactions", the linkers and caps were redesigned. The position of the affinity tags were changed in three constructs:

(i) H-CTPR3-Gp^N to CTPR3-Gp^N-H

(ii) H-CBD-Imp^C-CTPR3-Gp^N to CBD-Imp^C-CTPR3-Gp^N-H

(iii) H-CBD-Gp^C-CTPR3-Imp^N to H-Gp^C-CTPR3-Imp^N-CBD

Thus, the reaction scheme should proceed as per Figure 4.13.

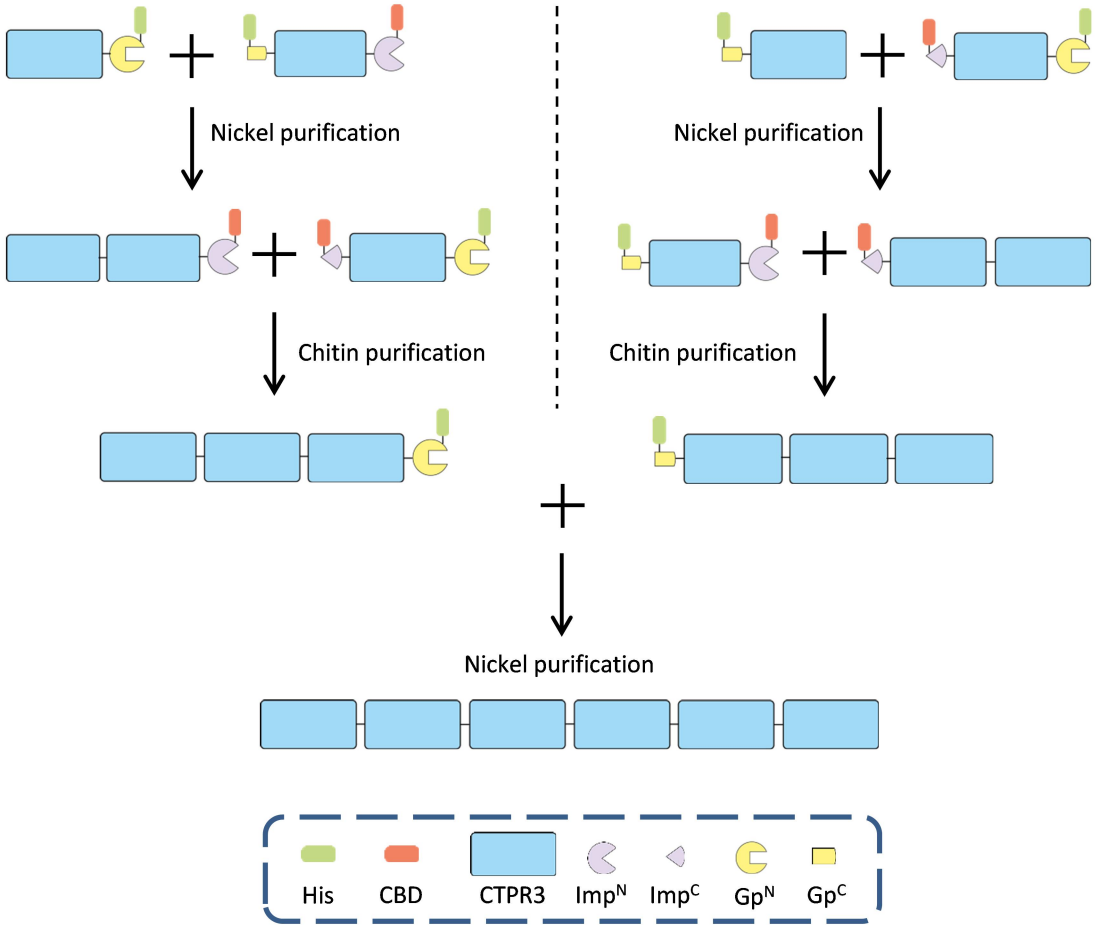


Figure 4.13 Flow chart of the process of making CTPR18. After each round of ligation reaction, the desired products were purified via affinity chromatography. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

With these changes, after each round of ligation, the product was purified by using either nickel or chitin affinity chromatography. Moreover, no concentration steps were performed until after the final round of ligation. This increased yields after each stage and did not reduce reaction efficiency.

4.6.2.1 First round of ligation and purification

The first round of extension for both caps was mediated by Gp split-inteins under the same reaction conditions as the first generation designs (50 μM concentrations, 50 mM Tris pH 8, 300 mM NaCl, 1 M Urea, 2 mM DTT) in a 1 to 1 ratio (Figure 4.14). Ni affinity chromatography was performed after the reaction to obtain the purified extended caps. From analysing the band intensity on the SDS-PAGE, the yield of each reaction achieved >85 % yield and the purification of the extended caps, CTPR6-Imp^{N} and CBD-Imp^{C}-CTPR6 were successful with only ~10 % loss of product.

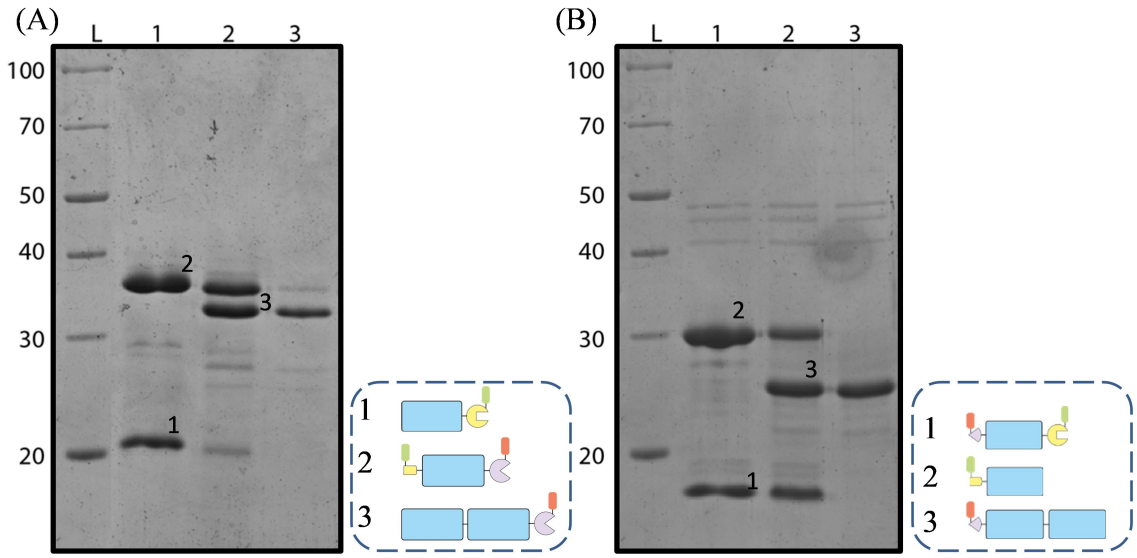


Figure 4.14 SDS-PAGE of the first round of reaction and purification. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time 0; Lane 2, after 1.5 hrs of reaction; Lane 3, purified products **(A)** CTPR6-Imp^{N}-CBD (44.2 kDa) and **(B)** CBD-Imp^{C}-CTPR6 (37.5 kDa) via Nickel affinity chromatography. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.6.2.2 Second round of ligation and purification

The second round of extension was mediated by Imp split-inteins (Figure 4.15). The conditions were identical to step 1 except a reaction concentration of 20 μM was used (as, the purified step 1 product was reacted without concentrating to reduce losses). The yield of the C-terminal reaction achieved ~85 %. Unfortunately, the reaction yield of the N-terminal reaction cannot be quantified because the product, CTPR9-Gp^{N}-H, and the linker, CBD-Imp^{C}-CTPR3-Gp^{N}-H, migrate at the same size on the denaturing gel. The purifications of the CTPR9-Gp^{N}-H and the H-CBD-CTPR9 via chitin affinity chromatography were successful with ~10 % loss of the product.

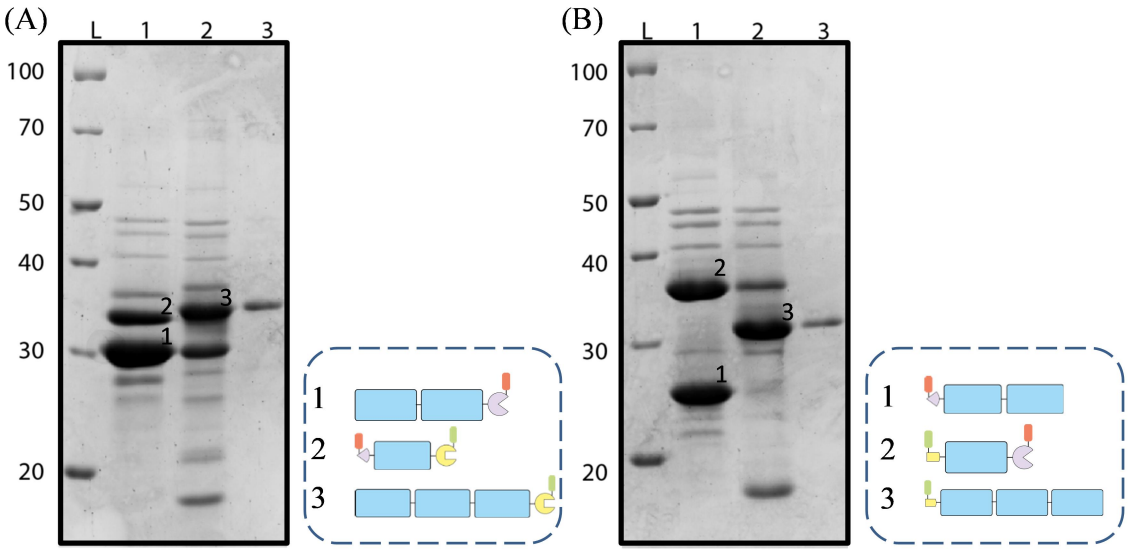


Figure 4.15 SDS-PAGE of the second round of reaction and purification. **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time=0; Lane 2, 3 hrs after the reaction; Lane 3, purified products **(A)** CTPR9-Gp^{N}-His (52.9 kDa) and **(B)** His-Gp^{C}-CTPR9 (47.7 kDa) via Ni affinity chromatography. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.6.2.3 Final ligation and purification

Finally, the purified CTPR9-Gp^{N}-H and H-CBD-CTPR9 were reacted in 1 to 1 ratio for 2 hrs in the same conditions, except the concentration was 6 μM (again to avoid losses during concentration). The reaction time was increased to ensure complete ligation. Nickel affinity chromatography was performed to remove the unreacted proteins (Figure 4.16). The yield of the ligation attained > 75 % with the purification step again only reducing the yield by 10 %. Figure 4.17 shows the stepwise assembly of the CTPR18.

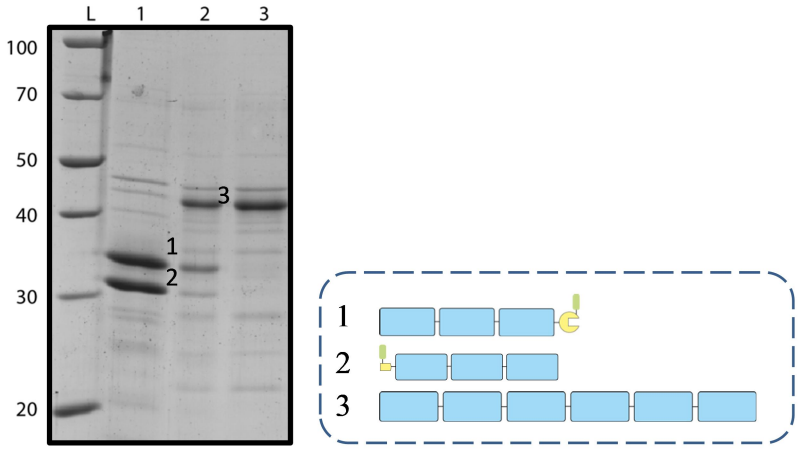


Figure 4.16 SDS-PAGE of the final round of reaction and purification. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time=0; Lane 2, 2 hrs after the reaction of CTPR9-Gp^{C}-H and H-CBD-Gp^{C}-CTPR9; Lane 3, purified product CTPR18 (81.5 kDa) via Nickel purification. Green represents 6-Histidine tag; blue represents CTPR3; and yellow represents Gp split-intein.

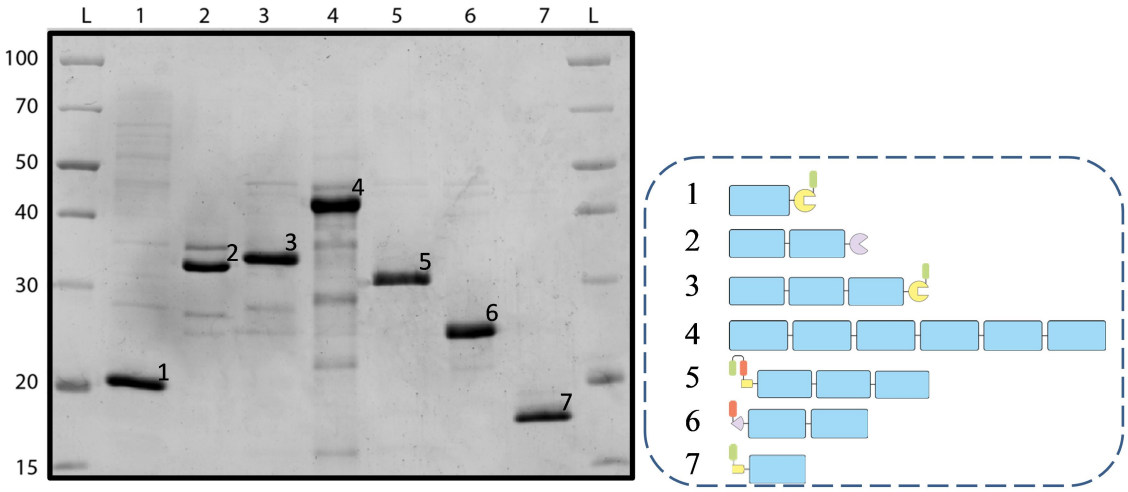


Figure 4.17 SDS-PAGE of the stepwise assembly of CTPR18. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, CTPR3-Gp^{N}-H; Lane 2, CTPR6-Imp^{N}-CBD; Lane 3, CTPR9-Gp^{N}-H; Lane 4, CTPR18; Lane 5, H-Gp^{C}-CTPR9; Lane 6, CBD-Imp^{C}-CTPR6; Lane 7, H-Gp^{C}-CTPR3. Green represents 6-Histidine tag; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

4.6.2.4 Secondary structure analysis of the assembled CTPR18

Far-UV CD spectroscopy analysis was performed to observe the secondary structure of the purified CTPR18 in solution, as described in Section 0. The purified CTPR18 fibril product was compared to a CTPR3 without the solvating helix (purified by Dr. C. Millership) that contains no split-inteins or affinity tags. The resulting CD spectra was calculated and plotted as shown in Figure 4.18. The far-UV CD spectra of the CTPR18 show that: (i) they are highly alpha-helical, and, importantly, (ii) have exactly six times the molar ellipticity in deg cm^2 $dmol^{-1}$ at 222 nm as that of the CTPR3 (Figure 4.18A). Figure 4.18B shows the molar ellipticity in deg cm^2 $dmol^{-1}$ $residue^{-1}$. The slight differences in the intensity at 208-235 nm are due to the extra amino acids (required for NCL and cloning) present in the CTPR18, which results in the number of residues of CTPR18 being 6.7 times more than the number of residues of CTPR3. The molar ellipticity deg cm^2 $dmol^{-1}$ $CTPR^{-1}$ (Figure 4.18C) shows that the extra amino acids present in CTPR18 did not affect the secondary structure of the CTPR18. Moreover, the ligation reaction had not caused any local unfolding on the secondary structure of the CTPR proteins.

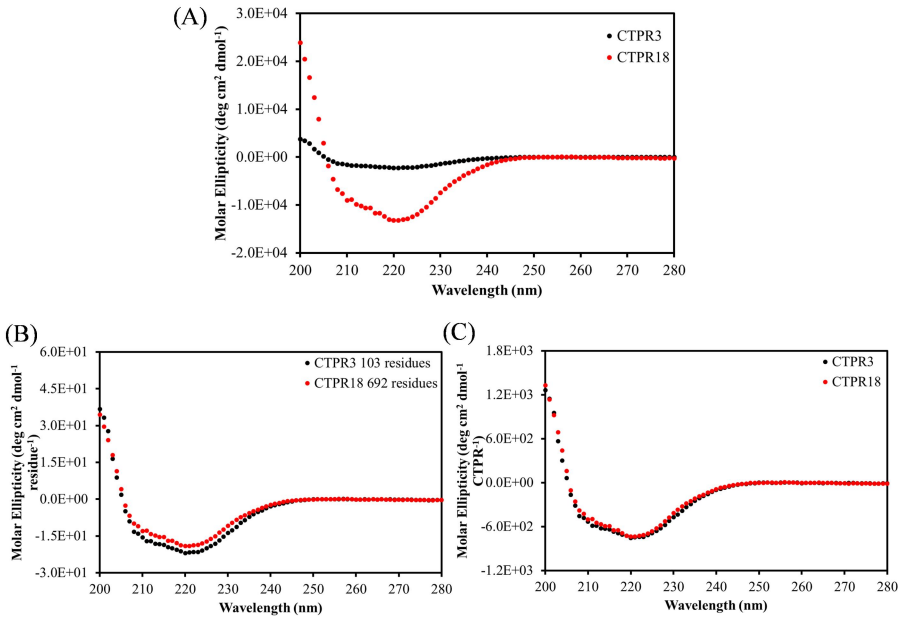


Figure 4.18 Far UV-CD spectra of CTPR18 (red) in comparison to a CTPR3 without split-inteins (black) with **(A)** molar ellipticity in deg cm^2 $dmol^{-1}$; **(B)** molar ellipticity in deg cm^2 $dmol^{-1}$ $residue^{-1}$; and **(C)** molar ellipticity in deg cm^2 $dmol^{-1}$ $CTPR^{-1}$.

4.6.2.5 Summary

The 2nd generation stepwise solution extension system successfully produced pure CTPR18. Moreover, the change of position of the affinity tags made the purification of

product after each round of reaction easier. Purification after each reaction eliminated any unwanted reactants and hence increased reaction yield. All of the reaction yields were in excess of 75 % and in certain cases as high as 90 %. Moreover, designing the system so that the product did not bind to the affinity resin and removing the need for concentrating after each round of purification, reduced losses to only 10 %. Thus, after each reaction and purification yields were on average 78 % yield with the final ligation and purification slightly lower at 68 %. Thus, the convergent 3-steps produced a final yield of the product CTPR18 of approximately 55 %.

4.7 Conclusion

In this chapter, it has been shown that the stepwise fibre extension mediated by this split-intein system is a much faster and higher yielding alternative to the MxGA intein system. Moreover, the split-intein system requires no activation steps. Both linker-tethered and solution syntheses were trialled and optimised. These showed:

Linker tethered synthesis: Immobilisation of the reactant proteins on chitin resin did not affect the initial Gp or Imp-mediated ligations. However, after two rounds of ligation, a significantly lower reaction yield was obtained for Imp-mediated ligation. This is likely due to either aggregation or steric hindrance. To improve reaction yields, spacer proteins could be introduced increasing the distance between the resin and the reactive split-inteins and reducing aggregation/steric hindrance during ligation. The reduction in reaction yield and loss of product during concentrating steps limited the tethered linker synthesis to two rounds of extension.

Solution synthesis: The solution synthesis was modified to enable purification of product after each stepwise extension. Importantly, by removing any concentrating steps, 90 % of the product was recovered from each of these. Thus, the overall yield of each step dictated that 3 rounds of extension produced a total yield of 55 % (each individual step produced > 75 %). Significantly, as the system can be convergently extended from both the N and C termini, larger structures can be produced than extension from only one terminus, with the same high yield. In conclusion, for the same number of extensions as the Mxe GyrA intein system, the split-intein mediated convergent synthesis enables the generation of larger products (joining 6 modules versus 4 modules) with a significantly greater yield (55 % versus 13 %) in a third of the time.

5 Assembly of Larger Cages

5.1 Introduction

This chapter explores the stepwise assembly of larger cages via iterative NCL addition mediated by the split-inteins described in Chapter 3 and 4. To show proof of principle non-functionalisable larger cages were produced by reacting half cage caps with CTPR3 "Spacer" linker constructs. Once this was established, cages were functionalised by replacing the CTPR3 in the linker with CTPR390. The CTPR390 has a motif that binds to specific pentapeptide tag (Cortajarena et al. 2004; Grove et al. 2012). Therefore, cargo that is tagged with the pentapeptide tag can be loaded onto the functional cages.

5.1.1 System design

The stepwise assembly requires three components: 2 half cages and a linker construct. In a similar manner to the fibre assembly, the larger cage assembly used a step-wise extension system. Here, one half cage cap was first reacted with a linker construct to create an extended half cage. Then the cage was completed by reacting the extended half cage with a complimentary half cage cap (Figure 5.1). After each reaction, affinity chromatography was performed to purify the product. Functional cages were assembled using the same method, except the linker contained a CTPR390 binding module as opposed to a "spacer" CTPR3 (H-Gp^{C}-CTPR390-Imp^{N}-CBD).

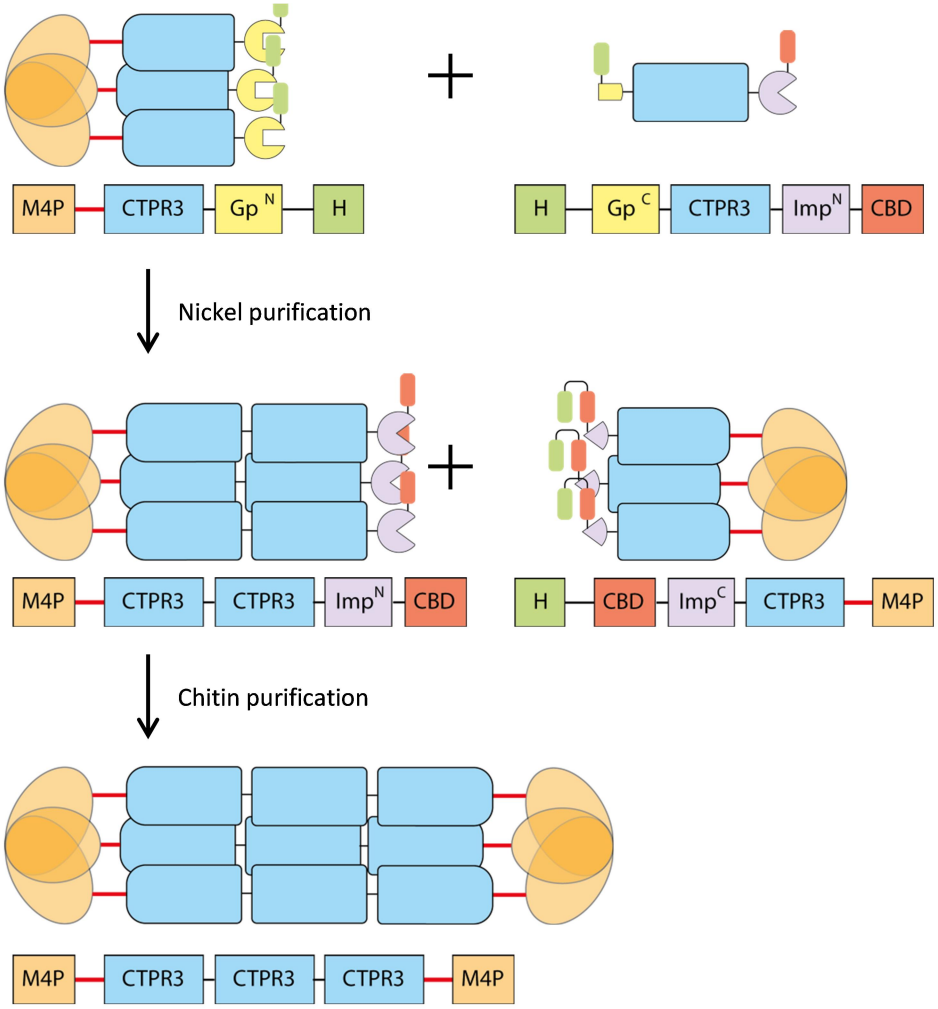


Figure 5.1 Schematic diagram of the process to assemble larger cages. First, the extended half cage cap was assembled by reacting M4P-CTPR3-Gp^{N}-H and H-Gp^{C}-CTPR3-Imp^{N}-CBD. The product was collected in the flow-through and wash fractions during purification. The extended half cage cap was ligated with H-CBD-Imp^{C}-CTPR3-M4P to make larger cages. The final product was purified via 2-step purification: affinity chromatography and SEC. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

5.1.2 Recombinant expression and purification of the required protein fusions

As seen in Figure 5.1, the three components required are M4P-CTPR3-Gp^{N}-H, H-Gp^{C}-CTPR-Imp^{N}-CBD and H-CBD-Imp^{C}-CTPR3-M4P. The expression and purification of the M4P-CTPR3-Gp^{N}-H and the H-CBD-Imp^{C}-CTPR3-M4P was described in Section 3.4.2, and the H-Gp^{C}-CTPR3-Imp^{N}-CBD was described in Section 4.4.2.

H-Gp^{C}-CTPR390-Imp^{N}-CBD was expressed and purified denatured (Section 2.3.3.2). It was refolded via step-down denaturant dialysis to determine its solubility condition. The final buffer condition was 50 mM Tris pH 7, 2 M urea, 300 mM NaCl, 5 mM DTT. The protein fusions were successfully refolded with 53.5 mg/L yield and high purity.

Samples were taken during the purification process and analysed by SDS-PAGE (Figure 5.2). As seen from the Figure 5.2, CTPR390 does not exhibit gel shift on SDS-PAGE as CTPR3 does.

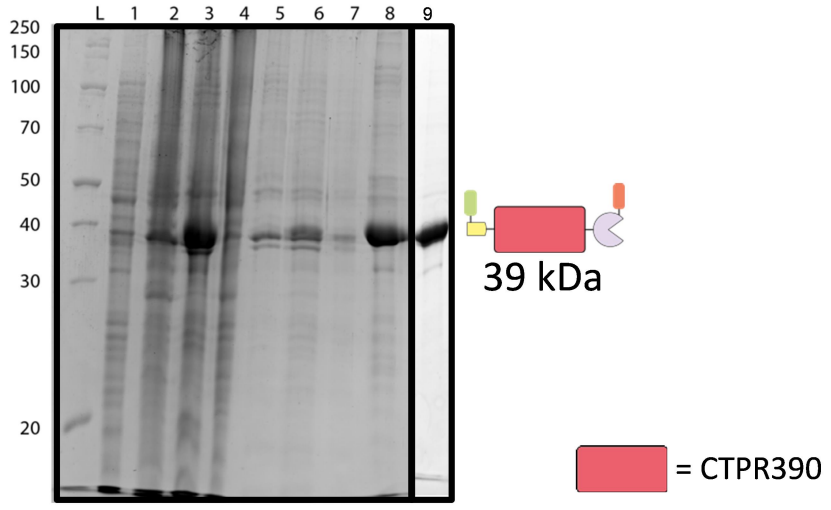


Figure 5.2 SDS-PAGE of the expression and purification of H-Gp^{C}-CTPR390-Imp^{N}-CBD. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, pre-induction culture; Lane 2, post-induction culture; Lane 3, denatured soluble lysate; Lane 4, denatured insoluble lysate; Lane 5, flow-through fraction; Lane 6-8, elution fractions; and Lane 9, purified dialysed H–Gp^{C}-CTPR390-Imp^{N}-CBD. Green represents 6-Histidine tag; pink represents CTPR390; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

5.2 Stepwise Assembly of larger cages

5.2.1 1st Step - NCL of half cage caps and linker

It is extremely important to obtain the highest yield possible for each reaction step in any multi-step reaction. It is particularly so here, given that each trimeric half cage requires each of its three sides to react. Therefore, both half-cage caps, with their differing split-inteins, were trialled in the first-step ligation to the linker fusion. This enabled the relative yields of a Gp and Imp mediated ligation to be assayed. In addition, differing excesses of linker fusion were also used to determine if this can increase the yield of extended half-cage product. 100 µL reaction volumes were used, with the half cage cap concentration always kept at 33 µM. The ratio of cap to linker was trialled at 1:3, 1:6 and 1:9, respectively, in standard reaction conditions, for 3 hrs, where the ratio is trimeric cage to monomeric linker. Figure 5.3 shows the results. The product yields were calculated against the amount of half cage caps left after the reaction.

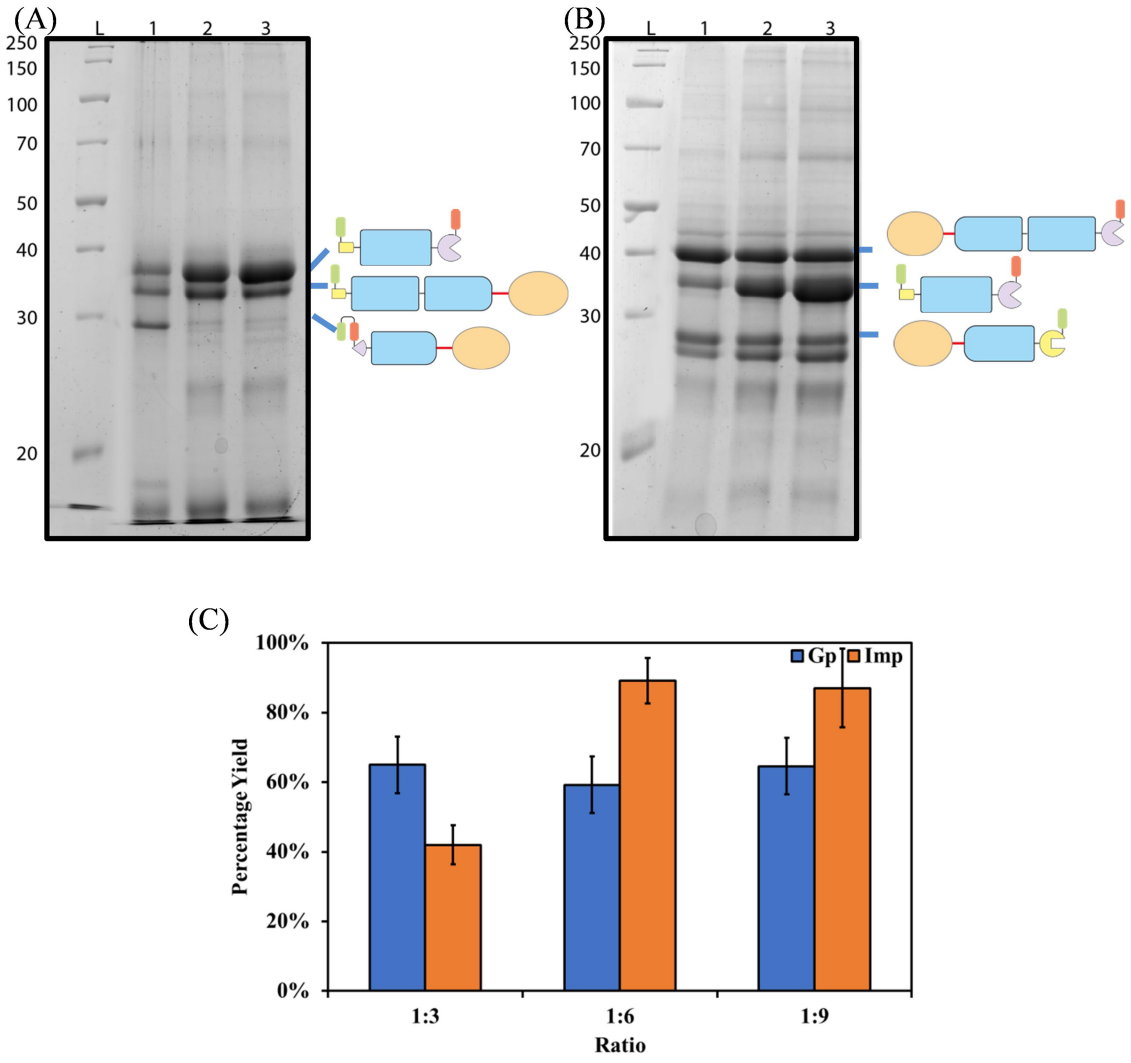


Figure 5.3. **(A and B)** SDS-PAGE of each reaction after 3 hrs. **(A)** Ligation reaction of H-CBD-Imp^{C}-CTPR3-M4P and H-Gp^{C}-CTPR-Imp^{N}-CBD in different ratios (product, H-Gp^{C}-CTPR6-M4P, 40 kDa). **(B)** Ligation reaction of M4P-CTPR3-Gp^{N}-H and H-CBD-Imp^{C}-CTPR3-M4P and H-Gp^{C}-CTPR3-Imp^{N}-CBD in different ratios (product, M4P-CTPR6-Imp^{N}-CBD, 50 kDa). **(A and B)** Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, 1 to 3 ratio; Lane 2, 1 to 6 ratio; and Lane 3, 1 to 9 ratio. **(C)** Calculated percentage yield of ligation mediated by the Imp and Gp split-inteins respectively, at different ratios. Error bars equate to standard deviation of multiple repeat experiments (at least twice in all cases). Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

The reaction yield mediated by the Gp split-intein appears to be independent of the reaction ratio. An average yield of 60 % was obtained for all three ratios. In contrast, the reaction yield mediated by the Imp split-intein increased substantially from 42 % when reacted in a 1:3 ratio to 90 % when an excess of linker was used (1:6 and 1:9). To determine the optimal time for the Imp split-intein extended cage reaction a time course of 3 hrs was recorded (Figure 5.4). As can be seen, the reaction was complete after 1 hr with ~90 % yield. Hence, to gain maximum yield of extended half cage product, the

Imp split-intein mediated reaction of half cage to linker was chosen. The reaction was carried out for 2 hrs and in a ratio of 1:6 (half cage cap trimer to monomeric linker).

Separation of Fully Ligated Product: Once reacted, the extended product was purified via chitin affinity chromatography (Figure 5.4A – Lane 9). Any unreacted half cage / linker fusions, partially spliced half cage caps and reacted split-inteins bound to the chitin resin, whereas the fully reacted extended half cages did not. The purified product was concentrated before proceeding with the next reaction. The yield of fully ligated extended half-cage after purification and concentration was approximately 50 %. Complete ligation and purity were confirmed by anti-CBD affinity tag Western blot (Figure 5.4B).

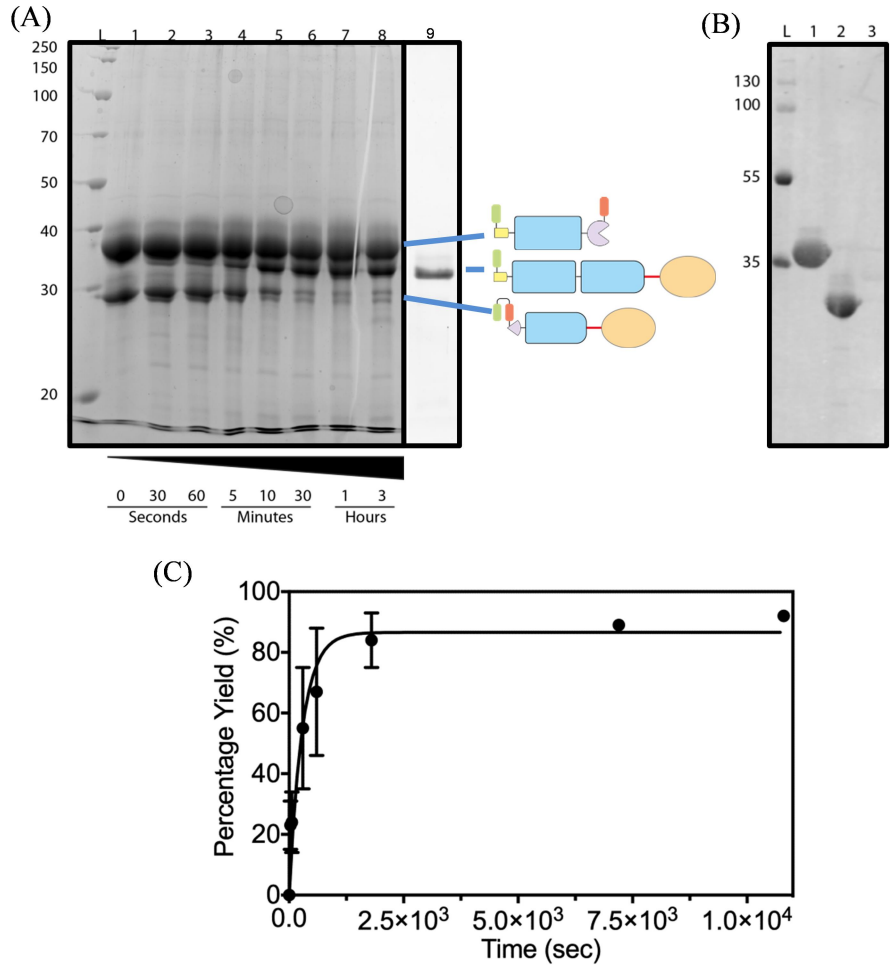


Figure 5.4 Ligation reaction of H-CBD-Imp^{C}-CTPR3-M4P and H-Gp^{C}-CTPR-Imp^{N}-CBD. **(A)** SDS-PAGE of the reaction in 1 to 6 ratio (H-CBD-Imp^{C}-CTPR3-M4P to H-Gp^{C}-CTPR3-Imp^{N}-CBD). Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time=0; Lane 2, 30 secs; Lane 3, 60 secs; Lane 4, 5 mins; Lane 5, 10 mins; Lane 6, 30 mins; Lane 7, 1 hr; Lane 8, 3 hrs; Lane 9, purified extended half cage (40 kDa). **(B)** Anti-CBD Western blot. Lane L, PageRuler Plus Prestained Protein Ladder in kDa; Lane 1, H-Gp^{C}-CTPR-Imp^{N}-CBD; Lane 2, H-CBD-Imp^{C}-CTPR3-M4P; Lane 3, H-Gp^{C}-CTPR6-M4P. **(C)** The initial rates of percentage ligated obtained from SDS-PAGE analysis fitted with a single exponential plus linear drift [(A*(1 -exp(-k*t))]. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; purple represents Imp split-intein; and red represents CBD.

5.2.2 2nd Step - Cage Closure

The next stepwise ligation reacted the extended half cage with a compatible Gp-half cage cap to form the closed cage. To favour the formation of discrete cages rather than networks, the cage closure was carried out with equimolar concentrations of reactants at 1 μM (as per Chapter 3 Gp-mediated cage synthesis). Excitingly, within 3 hrs, the percentage yield reached 70 %. This is consistent with the percentage yield of the smaller cage formation mediated by the Gp split-intein (Figure 5.5).

Separation of Fully Ligated Product: The fully ligated product was purified, as previously, by nickel chromatography and complete ligation confirmed by anti-His Western blot (Figure 5.5A-Lane 13 and B). Again, the percentage of complete ligation *versus* partial ligation was ~ 66 % (Figure 5.5D), which was consistent with the Gp-mediated cage formation.

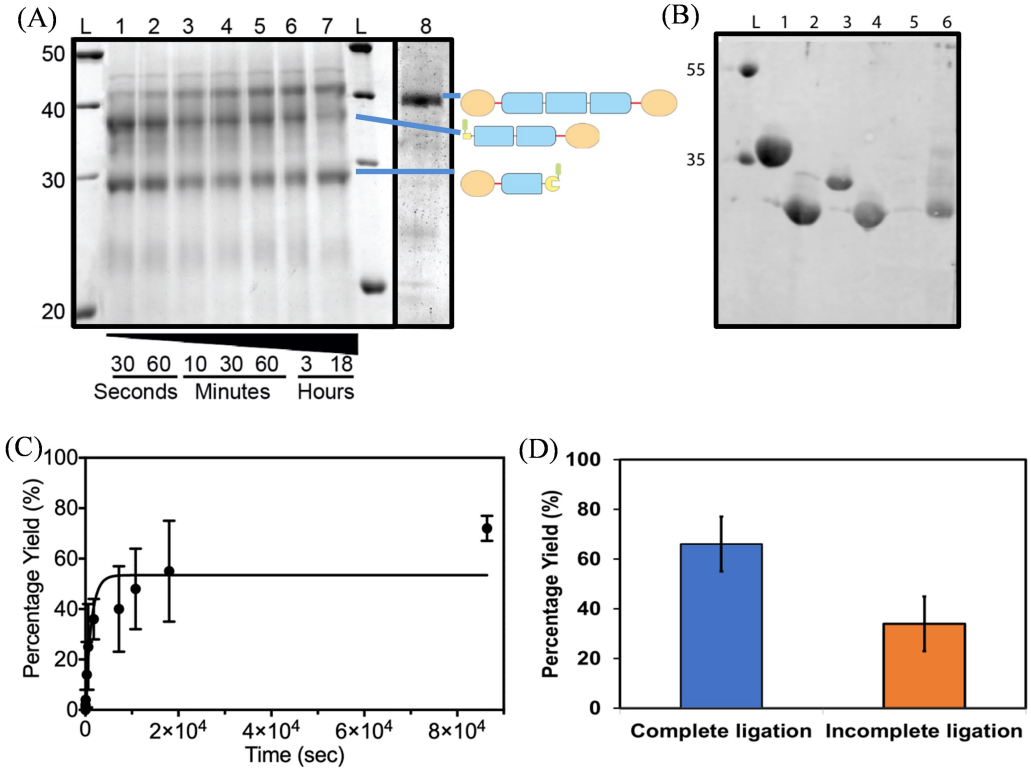


Figure 5.5 Ligation reaction of M4P-CTPR3-Gp^{N}-H and H-Gp^{C}-CTPR6-M4P. **(A)** SDS-PAGE of the reaction. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, 30 secs; Lane 2, 60 secs; Lane 3, 10 mins; Lane 4, 30 mins; Lane 5, 1 hr; Lane 6, 3 hrs; Lane 7, 18 hrs; and Lane 8, The purified product, M4P-CTPR9-M4P (51 kDa). **(B)** Anti-His Western blot. Lane L, PageRuler Plus Prestained Protein Ladder in kDa; Lane 1, H-Gp^{C}-CTPR-Imp^{N}-CBD; Lane 2, H-CBD-Imp^{C}-CTPR3-M4P; Lane 3, H-Gp^{C}-CTPR6-M4P; Lane 4, M4P-CTPR3-Gp^{N}-H; Lane 5, purified M4P-CTPR9-M4P; and Lane 6, elution fractions from the purification of M4P-CTPR9-M4P. **(C)** The initial rates of percentage ligated obtained from SDS PAGE gels analysis fitted with a single exponential plus linear drift [(A*(1 -exp(-k*t))]. **(D)** Calculated percentage of complete and incomplete ligation. Error bars equate to standard deviation of multiple repeat experiments (at least three in all cases). Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; and yellow represents Gp split-intein.

Separation of discrete cages from networks: As for the one-pot cages (Chapter 3), discrete cage structures were separated from extended networks via SEC. Figure 5.6A shows the results when a Superdex 200 10/30 column was used in standard reaction buffer. A monodispersed peak can be observed with an elution volume of 11.4 mL. Thus, the cage closure was successful, with a substantial proportion of the fully ligated product forming discrete assemblies rather than extended networks. When the elution volume of the extended cage was compared to that obtained for the two component cages, a difference of 0.5 mL was observed (extended cages eluted at 11.4 mL, and the two-component cage eluted at 11.9 mL). Fitting of the peak maxima of the larger cage to calibration standards, gave a molecular weight of 127 kDa (a difference of 20 % to the calculated trimeric molecular weight of 153 kDa) (Figure 5.6B). Thus, as expected, the extended cages form a slightly larger structure than the two-component cage, but not so large as to indicate a dramatic change in conformation.

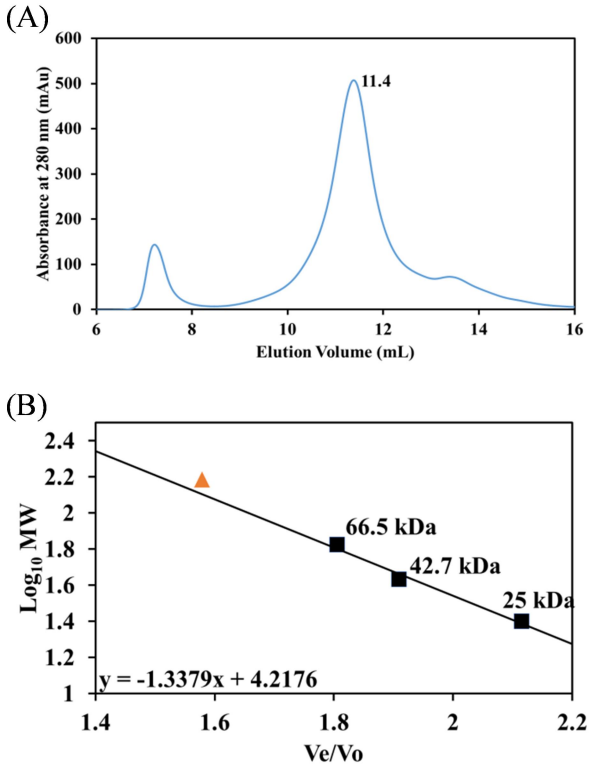


Figure 5.6 Trimeric analysis of the larger cages**. (A)** Superdex 200 10/30 SEC analysis. **(B)** The extended cages Ve/Vo (orange triangle) (elution volume/column void volume) of the extended cage plotted against their Log_{10} molecular weight on a standard curve. Black squares represent protein standards.

5.2.3 Structural analysis by SEC-SAXS

In a similar manner to Chapter 3, the extended cages were analysed by SEC-SAXS to obtain more information on their 3D shape. The collected SAXS data was processed and analysed as described in Section 2.5.7.1. Interestingly, initial inspection of the SAXS profile (Figure 5.7) showed differing features to the smaller one-pot cages described in Chapter 3. In particular, the region from 0.1 to 0.2 1/Å did not exhibit as prominent a concave feature. Thus, the addition of an extra 3 CTPR repeats has changed some features of the cage. This was to be expected, as the shape of the CTPR superhelix would be, at the very least, in a differing register with the addition of the extra CTPR motifs. Importantly, the curve still possesses several features that enable shape determination to a higher resolution.

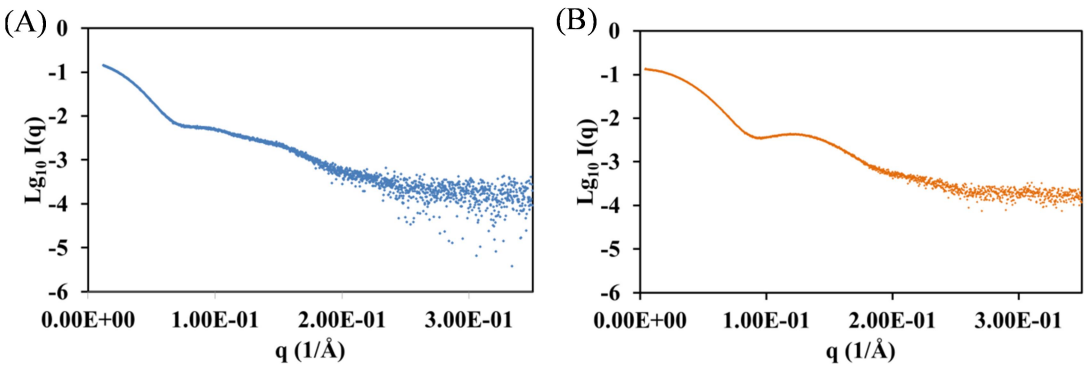


Figure 5.7 The SAXS profile of **(A)** larger cage product and **(B)** smaller cage product.

Initial analysis: Guinier and Kratky plot analysis of the SAXS data confirmed that the purified cages were monodisperse, rigid and multi-domained proteins. However, the Kratky plot analysis showed they are not as rigid as the smaller cages. The analysis showed that the extended cages are non-spherical and elongated with a radius of gyration (Rg) of 5 nm and a maximum linear particle diameter (Dmax) of 17.1 nm (Figure 5.8). Thus, the extended cages also have an elongated shape similar to the smaller cages, yet are slightly larger in both dimensions. The molecular weight of the extended cages obtained from the SAXS data is in agreement with that calculated from its amino acid sequence (157 kDa versus 155 kDa, respectively). Table 5.1 summarises the SAXS parameters of cage products obtained from the analysis and compares them to the smaller cages.

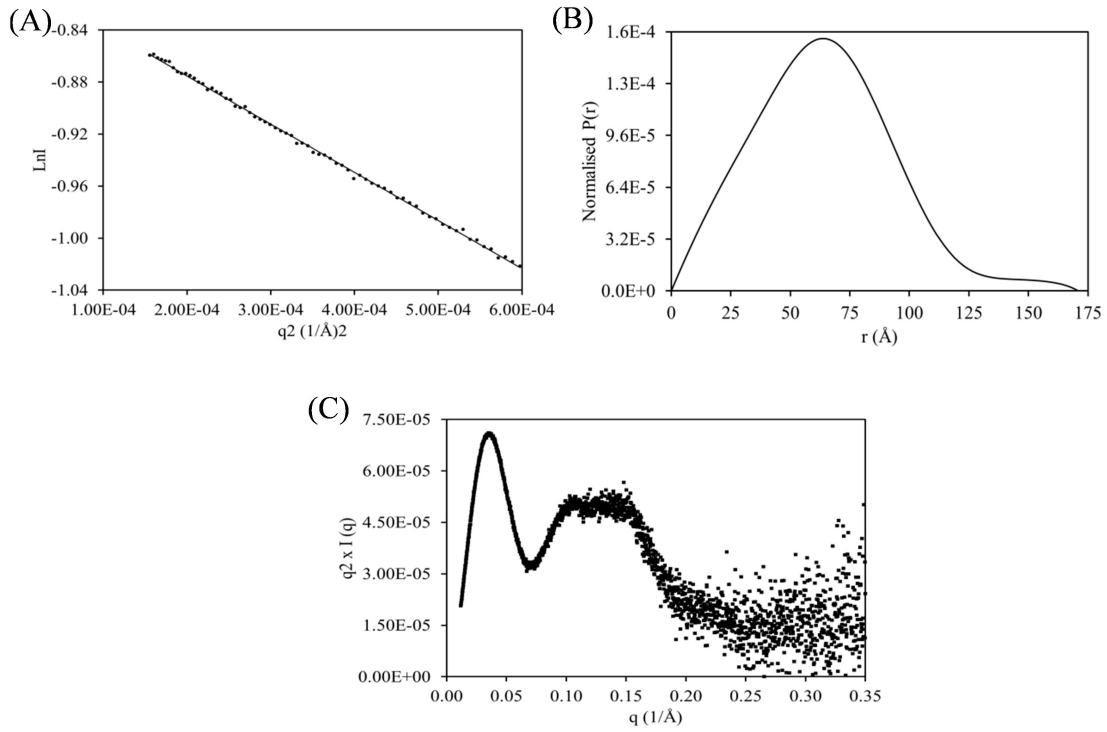


Figure 5.8 Analysis of SAXS of ligated cage and Kratky analysis of half-cage cap. **(A)** Guiner analysis, **(B)** distance distribution functions P(r) and **(C)** Kratky analysis of the SAXS for the ligated larger cages.

Table 5.1 SAXS parameters obtained from analysis of the purified extended cage and cage products

SAXS parameters	SAXS extended cage products	SAXS cage products
q range ($Å^{-1}$)	0.010 to 0.350	0.004 to 0.350
I(*0*) (Å)	0.160 +/-0.0002	0.134 +/-0.00011
*R*g (nm) (from Guinier)	5.00 +/-0.033	3.85 +/-0.023
*R*g (nm) (from *P*(*r*))	4.98 +/-0.006	3.82 +/-0.014
*D*max (nm) (from *P*(*r*))	17.1	12.6
Porod Exponent	3.4	3.7
MW^{SAXS} (Da)	157,050	110,568
$MW^{sequence}$ (Da)	154,533	113,369

Particle Reconstruction: Currently full particle reconstruction is still in progress. As with the smaller cages both *ab initio* and manual modelling construction techniques are being used. Below the results to date are summarised:

Ab initio – In a similar manner to the smaller cages, *ab initio* modelling was used with the program DAMMIF and P32-symmetry restraints. GASBOR was trialled, but failed, to generate sensible models. 60 models were generated and those that supported the biophysical data (5 models) were selected for averaging using DAMMAVER. An *ab initio* model (Figure 5.9A) was generated via DAMMIN using the averaged model as a template (as descript in Section 2.5.7.2). It generated an excellent fit to the SAXS curve with a χ^2 of 1.1 (Figure 5.9B). As can be seen, the model possesses a cage like structure, which is elongated by the addition of the TPR domains to the CTPR cage sides. It possesses an enclosed bipyramidal shape. However, the side domains do not seem as well defined as with the smaller cages.

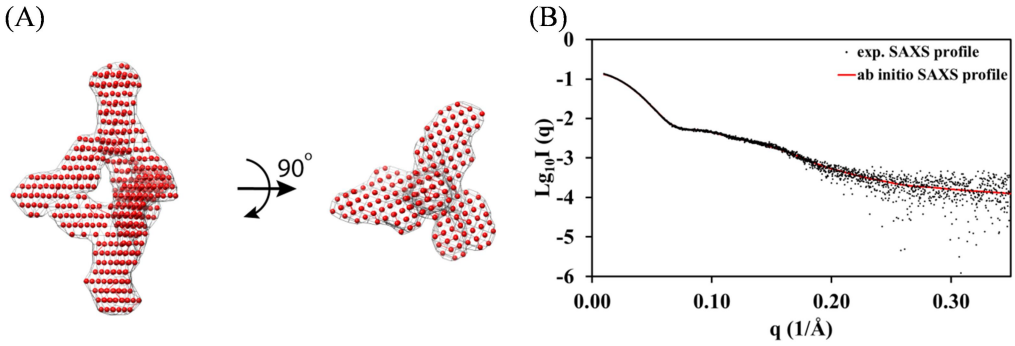


Figure 5.9 **(A)** Two orientations of the *ab initio* DAMMIN generated model. **(B)** Experimental SAXS profile (black circles) of ligated extended cages overlaid the *ab initio* DAMMIN generated model SAXS profile (red line).

Manually generated atomic model – Models were generated as described in Section2.5.7.3. To date, we have generated several models. We have trialled: (1) a trigonal pyramidal model that uses a docked CTPR9 superhelix, (2) a trigonal pyramidal model with a docked CTPR6 superhelix attached to an undocked CTPR3 and (3) a trigonal pyramidal model with all CTPR3 sides undocked. Models (1) and (3) gave a χ^2 of >50. This is in agreement with the Kratky plot analysis, where the extended cage is only slightly more flexible than smaller cages. Whereas, model (2) gave a χ^2 of 2.85 (Figure 5.10). It is expected that the CTPR6 formed in the first step will maintain the CTPR6 superhelix structure but the third CTPR3 from the second ligation does not dock into a superhelix. Presently, more differing models are being generated to better fit the SAXS profile.

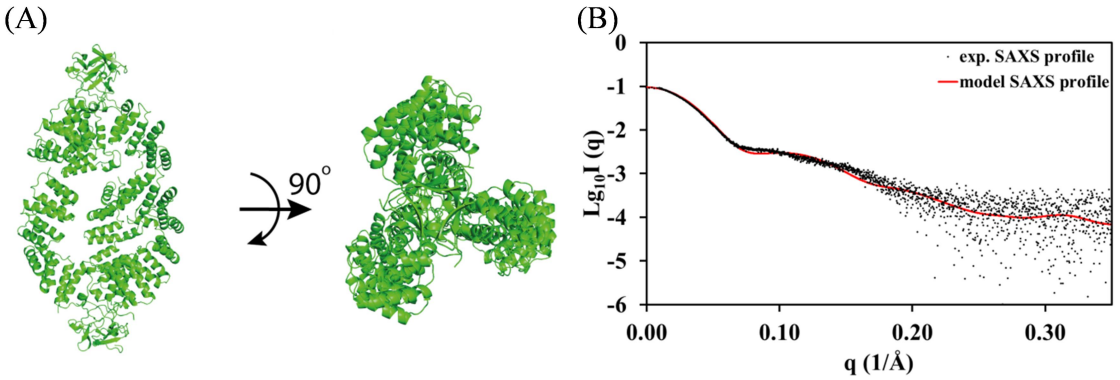


Figure 5.10 **(A)** Two orientations of the best atomic model generated. **(B)** Experimental SAXS profile (black circles) of ligated extended cages overlaid the generated model SAXS profile (red line).

5.2.4 Summary

We have shown that the stepwise assembly system successfully yielded ~23 % of highly pure extended cages, with a substantial proportion of the fully ligated product forming discrete assemblies rather than extended networks. Hence, this three-component system permits a realistic scalable extension limit of two reactions. However, if a convergent ligation strategy as described chapter 4 was used, this could enable the connection of 4 models together. Both SEC and SEC-SAXS, as expected, show that the extended cages form a slightly larger structure than the two-component cages, but not so large as to indicate a dramatic change in conformation. Model reconstruction of the larger cage is on-going.

5.3 Assembly of functional cages

As the stepwise assembly successfully produced larger cages, the next phase was to form functional cages. There are two ways of functionalising the cages: (1) Replace the CTPR3 in the half cage cap constructs with a binding module and perform a two-component synthesis; or (2) use a three-component two-step synthesis and use a linker with a binding module. Here, the second approach was explored.

For the binding module, we used CTPR390. This is a CTPR3 variant whose pentapeptide binding pocket has been designed to recognise the -DESVD sequence (Speltz, Nathan, and Regan 2015). Thus, in future, molecules of interest can be tagged with the pentapeptide sequence and then 'loaded' onto the nanostructure.

5.3.1 NCL of half cages and functional linker

Given the success of the Imp-mediated half-cage extension ligation, it was decided to form an extended half cage by reacting the Imp^C tagged half cage cap with a CTPR390 containing linker construct. The reaction was conducted as previously, *i.e.* in the ratio of 1 to 6 (half cage cap to linker). Figure 5.11 shows the reaction of H-CBD-Imp^C-CTPR3-M4P and H–Gp^C-CTPR390-Imp^N-CBD after an 1 hr incubation. The ligation reaction was in 50 mM Tris pH 7, 2 M urea, 300 mM Tris, 5 mM DTT. 2 M urea was used as the linker was not soluble in 1 M Urea. During the 1 hr reaction, a large quantity of precipitation was observed. Fortunately, the precipitant could be denatured in 6 M GuHCl and refolded into reaction buffer supplemented with 3 M urea. A SDS-PAGE of the reaction, the precipitant, refolded precipitant and soluble reaction product is shown in Figure 5.11. Unfortunately, the CTPR390 linker and extended half cage product migrated at the same size on the denaturing gel. Both the refolded and soluble reaction product was subjected to chitin affinity chromatography. As per the previous extended half cage reaction, every protein will bind to the resin except the fully ligated extended half cage product. The purified, fully ligated, product is shown in Figure 5.11A-Lane 5. Optimisation of the purification steps (using new chitin resin and slower flow rates) increased the overall yield from 20 % to 50 %.

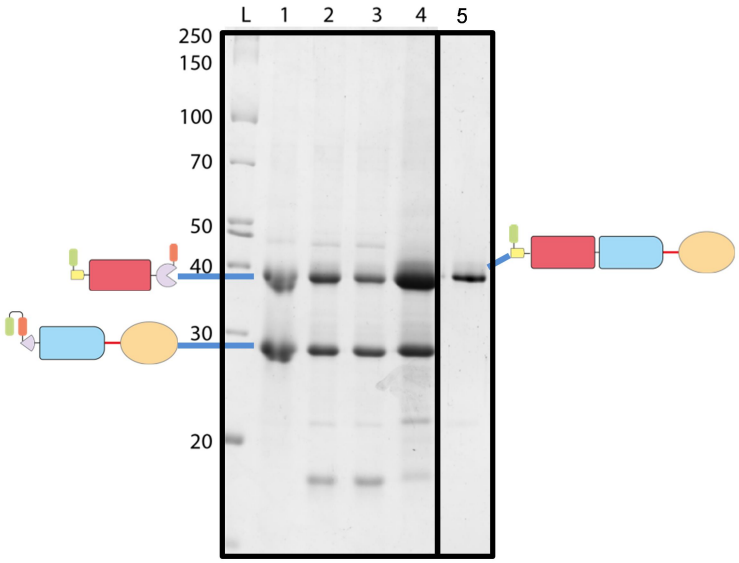


Figure 5.11 SDS-PAGE of the Imp-mediated ligation to form an extended half cage. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time=0; Lane 2, 1 hr of reaction; Lane 3, insoluble fraction of the reaction; Lane 4, refolded reaction mixture; and Lane 5, purified product via chitin chromatography. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; pink represents CTPR390; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

5.3.2 Functional Cage Closure

The cage was closed with Gp mediated ligation of M4P-CTPR3-Gp^N-H cap and the extended functional half cage cap. Similar to section 5.2.2, H-Gp^C-CTPR390-CTPR3-M4P and M4P-CTPR3-Gp^N-H were reacted in 1 µM to favour the formation of discrete cages. The reaction was left for 5 hrs in reaction buffer supplemented with 3 M urea. Figure 5.12A shows the reaction at different time points over 5 hrs. The reaction was completed within 3 hrs with 60 % yield. This is in line with both the smaller and extended Gp-mediated cage reactions (Figure 5.12B).

Separation of Fully Ligated Product: The ligation products, unreacted starting material and split-inteins were then separated via Ni affinity chromatography. As with the smaller cages, unreacted and partially reacted proteins will bind to the Ni, whereas fully ligated proteins will not. Figure 5.11C compares those proteins bound to the resin with those that were not. This allowed the yield of complete to incomplete ligation products to be delineated. As can be seen, the percentage of complete ligation is much lower (37 %) compared to both the smaller and extended Gp-mediated cage reactions (66 %).

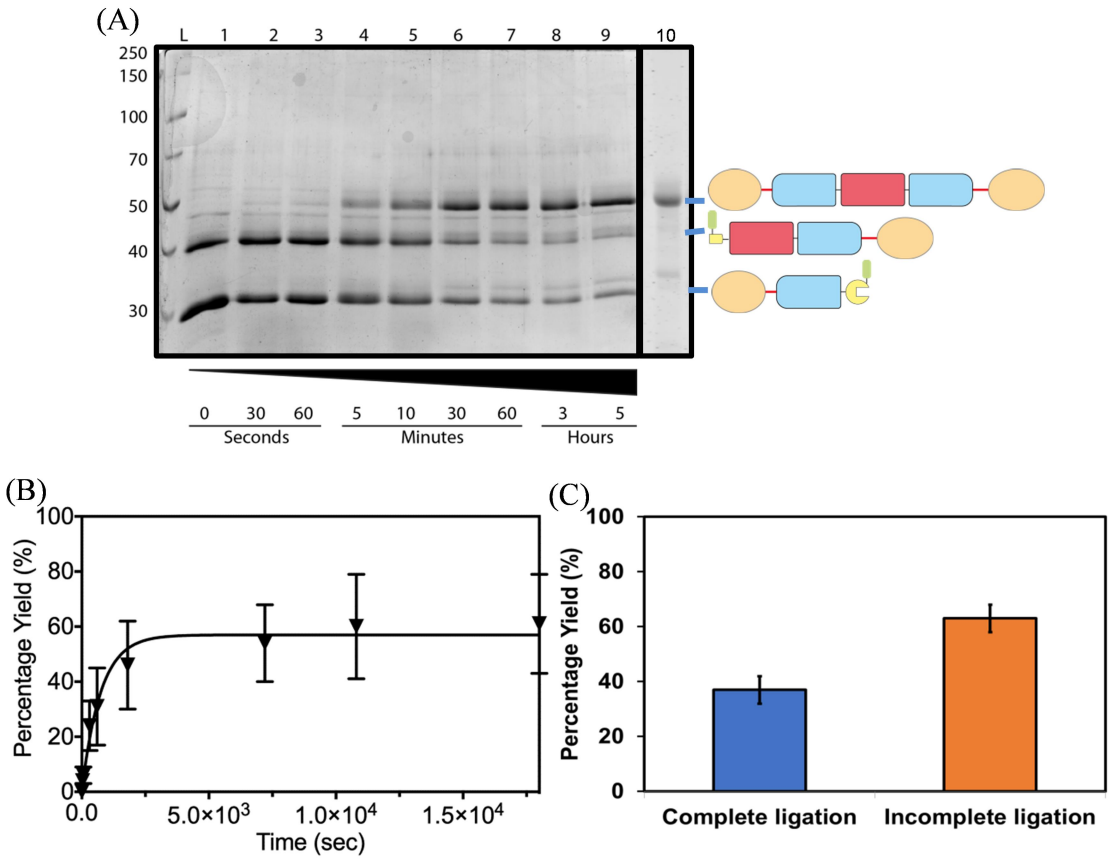


Figure 5.12 Gp-mediated ligation to produce extended cages. **(A)** SDS-PAGE of the reaction. Lane L, PageRuler Broad Range Unstained Protein Ladder in kDa; Lane 1, time=0; Lane 2, 30 secs; Lane 3, 60 secs; Lane 4, 5 mins; Lane 5, 10 mins; Lane 6, 30 mins; Lane 7, 60 mins; Lane 8, 3 hrs; Lane 9, 5 hrs; and Lane 10 purified of the product via Ni affinity chromatography (product formed, M4P-CTPR3-CTPR390-CTPR3-M4P, 51 kDa). **(B)** The initial rates of the percentage ligated obtained from SDS PAGE gels analysis fitted with a single exponential plus linear drift [(A*(1 -exp(-k*t))]. **(C)** Calculated percentage of complete ligation and incomplete ligation. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; pink represents CTPR390; purple represents Imp split-intein; red represents CBD; and yellow represents Gp split-intein.

Separation of discrete cages from networks: To separate discrete functional cages from networks and other impurities, SEC was performed (Figure 5.13). A Superose 6 10/300 column was run in 50 mM Tris pH 8, 3 M urea, 300 mM NaCl, 5 mM DTT buffer. A peak was observed at a similar size to the previously extended cage (15 mL in 50 mM Tris pH 8, 1 M urea, 300 mM NaCl, 5 mM DTT buffer – orange dotted line). The difference in elution volumes (1 mL) was likely due to the concentration of denaturant used in the buffer.

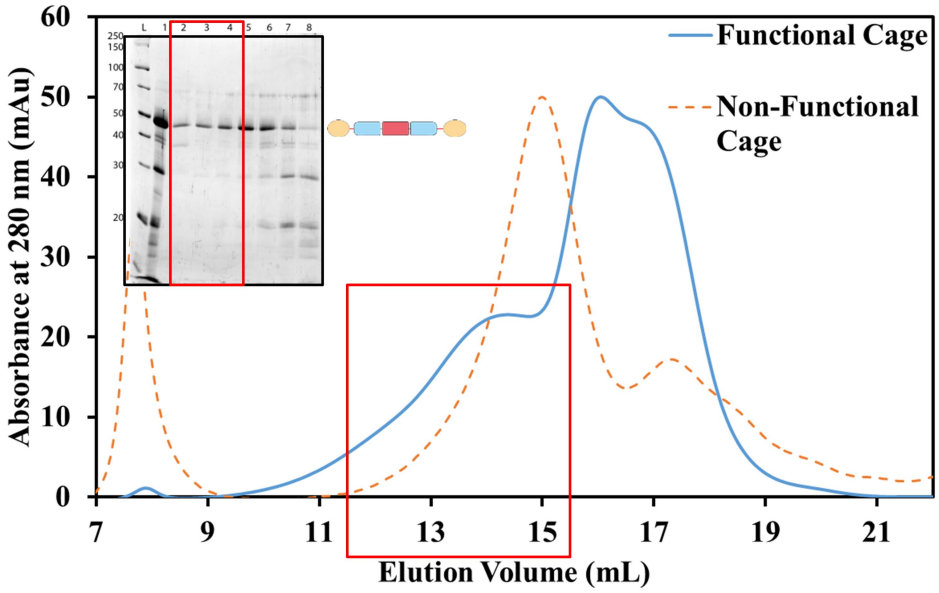


Figure 5.13 Purification of discrete functional cages via SEC (blue line). Superose 6 10/300 column was run in 3 M urea, 300 mM NaCl, 50 mM Tris pH 7, 5 mM DTT buffer. The possible discrete cages (red box) fractions were pooled and concentrated. Orange dotted line – larger cage in 50 mM Tris pH 8, 1 M urea, 300 mM NaCl, 5 mM DTT buffer. Green represents 6-Histidine tag; orange represents M4P; red represents linker; blue represents CTPR3; and pink represents CTPR390.

5.3.3 Summary

Functional product has been successfully assembled with final yield of 10 % using a 2-step assembly system. The product was effectively purified through affinity chromatography and SEC. Further characteristics of the functional product *i.e.* SEC-SAXS are required to confirm the structure formed.

5.4 Conclusion

In this chapter it has been shown that the designed protein fusions *i.e.* half cage caps and linkers, can be successfully utilised to assemble larger cages in a stepwise manner and can incorporate modules that have binding capacity. In particular, a two-step extension system was shown to form discrete cages and incorporate sides with either non-binding or binding modules. These could be purified to homogeneity with a combination of affinity and size exclusion chromatography.

Significantly, when the extended non-binding cages were analysed with SEC and SAXS, they were shown to produce elongated cage structures that were larger than the one-pot cages (Chapter 3), but not so large as to suggest a change in the expected trigonal bipyramidal structure. Moreover, the expected molecular weight agreed with that calculated from the amino acid sequence (157 kDa). Importantly, from the initial analysis, it can be concluded that the structure is a cage with P3-symmetry and an enclosed cavity. Further structure reconstruction is still in progress.

When the CTPR binding module was used in the linker construct, the yield of final cages was decreased. This might be due to the higher concentration of denaturant used or a more inherent property of the binding module (more aggregation prone). Once the cage closure reaction was concluded, the discrete cages could be purified to homogeneity and were found to be stable (non-aggregation prone) in 1 M urea. This is in contrast with non-functionalised cages that are stable in non-denaturing aqueous buffer. The increased aggregation potential is likely due to the instability of the binder module used. However, this may be overcome by using other, less aggregation-prone CTPR binding modules (Speltz, Nathan, and Regan 2015). Nonetheless, this chapter demonstrates functional cages can be produced with relative ease and thus provides a general route to enable the loading of cargo.

6 Conclusions

6.1 Discussion and further work

This thesis explored the use of genetically encoded NCL to control the assembly of designed and recombinantly produced protein fusions into specific user-defined nanostructures and biomaterials. Firstly, in Chapter 3, fusion proteins were designed that are capable of self-assembling into protein nanocages with a central hollow cavity. In Chapter 4, the limits of the NCL system were examined by determining the most NCL reactions that can be performed to form a linear structure. Finally, in Chapter 5, we further explored cage forming fusion proteins by designing in functionality.

6.1.1 Design and assembly of symmetric protein cages

In Chapter 3 and 5, we have proof that genetically programmed split-intein mediated NCL can be successfully employed to assemble modular proteins designed with simple geometric symmetry into user-defined protein cages. Two pairs of split-inteins, Imp and Gp split-inteins were used (i) separately to drive two-component assembly, assembling two half-cage modules (M4P vertex and CTPR3 sides) to form a cage; and (ii) together to drive three-component assembly, assembling two half cages and a linker to form larger cages. Under mild conditions, mixing compatible oligomeric protein fusions resulted in rapid and irreversible ligations with high yield of ligated products (via peptide bond formation). The fusion proteins were designed to enable discrete, fully ligated, cages to be easily and efficiently separated to homogeneity from each reaction in a 2-step purification. Significantly, this process generated the expected square bipyramidal structures ranging from 113 kDa (for the two-component, one reaction) to 157 kDa (three-component, two-step reactions). Furthermore, the cage produced from the two-component system was shown to contain a central hollow cavity that could accommodate cargo up to 70 × 55–60 Å. As for the larger cages, from the initial analysis of SEC-SAXS, it can be concluded that the structure is a cage with P3-symmetry, slightly flexible and has an enclosed cavity. Further structural reconstruction is still in progress. Interestingly, although both split-inteins gave high ligation yields, a

greater yield of fully ligated discrete cages was obtained when lower reactant concentrations and the slower reacting Imp split-intein were used.

The use of extendable CTPR sides additionally provided a method for loading cargo into the central hollow cavity of the nanostructures by using a binding module as the linker. However, when the CTPR binding module, CTPR390, was used in the linker construct, the yield of final cages was decreased. This was due to a combination of the higher concentration of denaturant used and the more aggregation prone nature of the binding module. Once the cage closure reaction was concluded, the discrete cages could be purified to homogeneity and were found to be stable in non-denaturing buffer for the non-functional cage and 1 M urea for the functional cage. The increased aggregation potential of the functional cage was due to the instability of the binder module used. Fortunately, this may be overcome by using other, less aggregation-prone CTPR binding modules (Speltz, Nathan, and Regan 2015). Moreover, each CTPR binding-modules recognises specific pentapeptide tag sequences (Speltz, Nathan, and Regan 2015). Thus, molecules of interest could be tagged with a pentapeptide sequence and then ‘‘loaded’’ onto the nanostructure via the binding module, creating a generic loadable system. The use of proteins such as TPRs, which are natural binding proteins, is an advantage of our assembly systems compared with others: for example, those based on coiled-coils that do not possess such intrinsic binding capabilities.

In conclusion, our protein assembly system provides a more general route to producing protein cages that avoids many time-consuming and system-specific processes (for example, those requiring computational design). No bioconjugation, chemical modification, or postligation refolding steps were required, and only a short sequence was inserted at the point of the NCL.

6.1.2 Limitation of the systems

In Chapter 4, we explored the maximum number of NCL that can be achieved via iterative fibre formation. A total of four fusion proteins were required: two linker fusion proteins, with flanking orthogonal intein pairs and two capping fusion proteins, each with a single complementary intein domain either on the N or C-terminus. Two methods of stepwise extension were investigated: (i) tethered linker system - whereby the growing product is eluted from immobilised fusion proteins via affinity tags and (ii) in solution synthesis - whereby protein fusions are reacted in solution and the product of

each step affinity purified. Finally, the limits of both syntheses were delineated (*i.e.* the number of extensions possible). Both syntheses concluded that 3 NCL reactions can be achieved in a short time with ~55 % final yield. Excitingly, as the system can be convergently extended from both the N and C termini, larger structures can be produced compared to extension via intein mediated NCL where it can only extended from the C-terminus (Harvey, Itzhaki, and Main 2018). In addition, linkers with binding modules, *i.e*. CTPR390, can be utilised to assemble functionalisable fibres.

6.2 Further directions

6.2.1 Creating new geometries

Nature has a vast range of protein domains that engineer different geometric shapes and coupled with these high-yielding split-inteins to drive the reaction, there are many opportunities for exploiting our self-assembling protein system. Alternative protein structures could be formed with differing vertices geometry. For example, replacing the trimeric M4P used for cage assembly with a hexameric PduA protein. In nature, PduA is a 37.4 kDa homohexamer that is used in the formation of bacterial microcompartments. Its selective permeable pore regulates the influx of substrate and efflux of toxic intermediate (C Chowdhury et al. 2015). Similar to M4P domain, both the N and C-termini of PduA are located on the same protein face (Figure 6.1A) (Pang et al. 2014). Using PduA as the vertices, linear repeat proteins as the sides and a split-intein driving force to close both ends, a hexameric rod could be assembled (Figure 6.1B). Furthermore, hexameric PduA packs to form flat sheet in the crystal or can self-assemble to from synthetic protein nanotubes in low salt concentration solution (Pang et al. 2014; Uddin et al. 2018). Hence, a bilayer sheet can be assembled with the hexameric rod in low salt concentration solution (Figure 6.1C), a biomaterial that could be loaded with biomolecules for biopharmaceutical approaches or used to filter heterogeneous particle solutions.

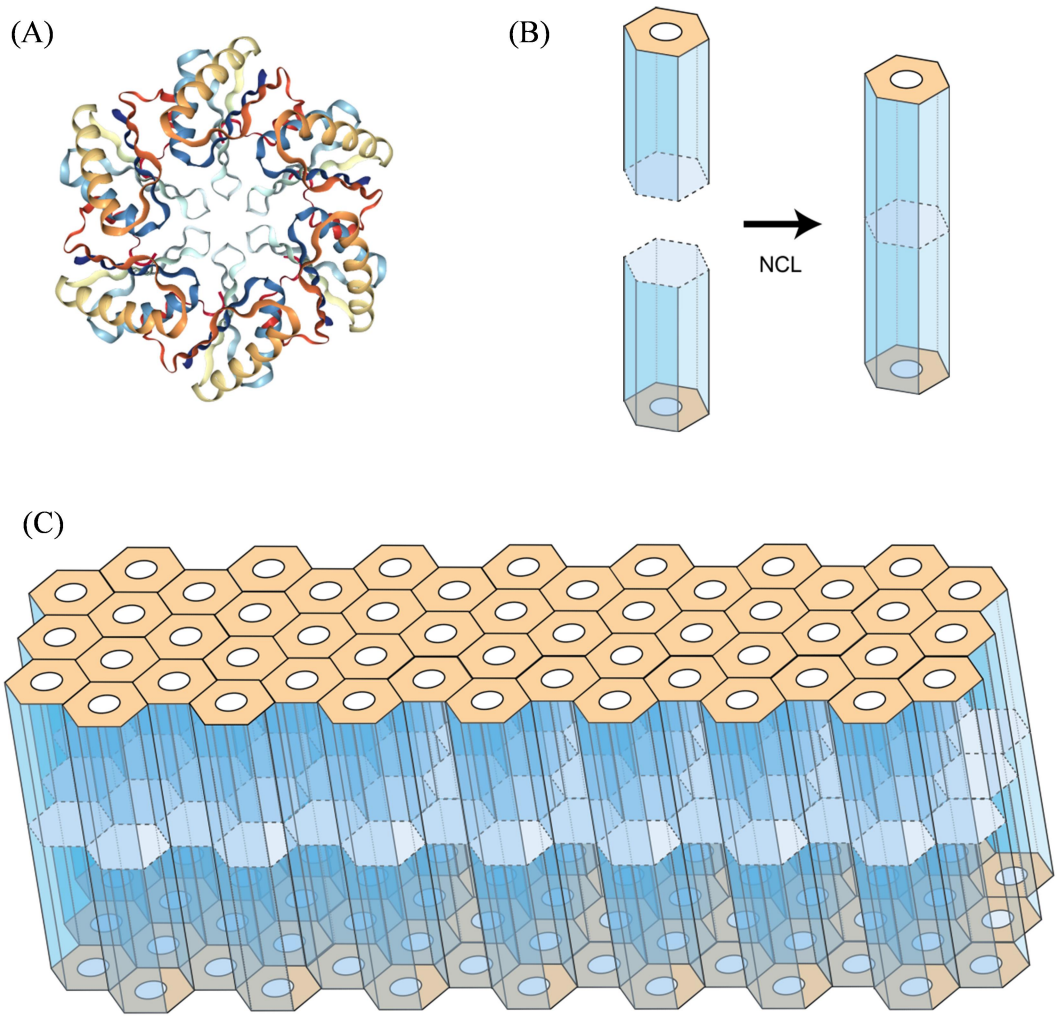


Figure 6.1 Assmbly of hexameric rod. **(A)** Crystal structure of PduA (PDB ID 3NGK). **(B)** Schematic diagram of the assembly of hexameric rod via split-intein mediated NCL. **(C)** Schematic diagram of the assembly of bilayer sheets. Orange represents PduA; and blue represents CTPR3.

6.2.2 Creating multifunctional nanocages

The one-pot cage and two-step cage assembly described in this thesis has potential biotechnological and biomedical applications, using specific modules to bind target load. For example, Speltz *et. al.* designed three TPR binding modules (TRAPs) that binds to its cognate peptide and exhibits low cross-reactivity with the peptides bound by the other TRAPs (Speltz, Nathan, and Regan 2015). Hence, the CTPR3 used in this thesis can be replaced with any of these TRAPs to assemble two or more-components nanocages (Figure 6.2A). Then, different cargos and/or site recognition peptides could be tagged with its cognate peptide and loaded onto the cages (Figure 6.2B). These multifunctional nanocages have potential use in targeted drug delivery where it allows medicine to be transported to an exact location in the body, *e.g.* a cancerous tumour, while minimising the damage to tissues surrounding the treatment site and degradation.

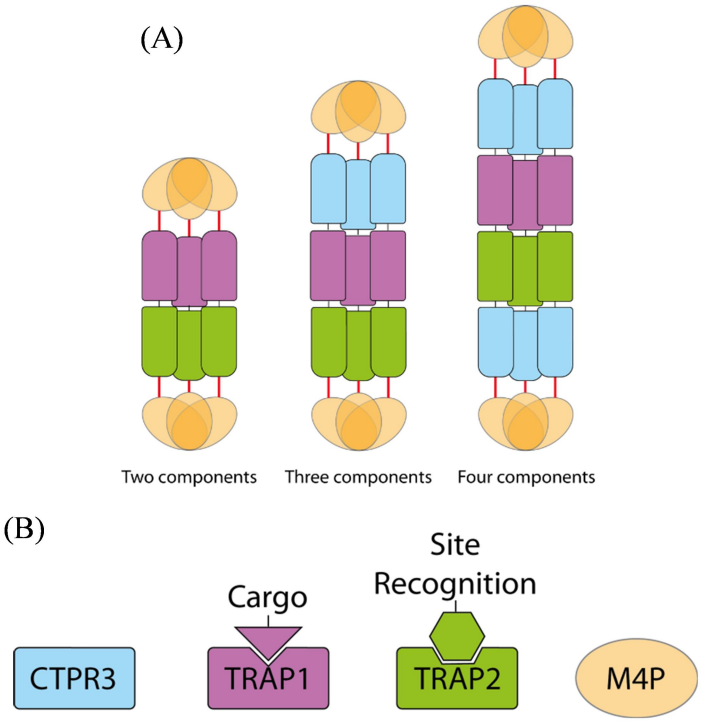


Figure 6.2 **(A)** Schematic diagram of the differing nanocages that could be assembled. **(B)** The domains used in assembling the nanocages and potential load that each TRAP can carry. Orange represents M4P; red represents linker; blue represents CTPR3; purple represents TRAP1; and green represents TRAP2.

7 References

"About RosettaCommons." 2015. https://www.rosettacommons.org/about.

Anwar, Rashid A. 1990. "Elastin: A Brief Review." *Biochemical Education* 18 (4): 162–66. https://doi.org/10.1016/0307-4412(90)90121-4.

Armstrong, C T, A L Boyle, E H Bromley, Z N Mahmoud, L Smith, A R Thomson, and D N Woolfson. 2009. "Rational Design of Peptide-Based Building Blocks for Nanoscience and Synthetic Biology." *Faraday Discuss* 143: 305–72. http://www.ncbi.nlm.nih.gov/pubmed/20334109.

Baldock, Clair, Andres F. Oberhauser, Liang Ma, Donna Lammie, Veronique Siegler, Suzanne M. Mithieux, Yidong Tu, et al. 2011. "Shape of Tropoelastin, the Highly Extensible Protein That Controls Human Tissue Elasticity." *Proceedings of the National Academy of Sciences of the United States of America* 108 (11): 4322. https://doi.org/10.1073/PNAS.1014280108.

Bale, J. B., S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, et al. 2016. "Accurate Design of Megadalton-Scale Two-Component Icosahedral Protein Complexes." *Science* 353 (6297): 389–94. https://doi.org/10.1126/science.aaf8818.

Baxevanis, A D, and C R Vinson. 1993. "Interactions of Coiled Coils in Transcription Factors: Where Is the Specificity?" *Curr Opin Genet Dev* 3 (2): 278–85. http://www.ncbi.nlm.nih.gov/pubmed/8504253.

Blaber, Michael, and Jihun Lee. 2012. "Designing Proteins from Simple Motifs: Opportunities in Top-Down Symmetric Deconstruction." *Current Opinion in Structural Biology* 22 (4): 442–50. https://doi.org/10.1016/j.sbi.2012.05.008.

Bragulla, Hermann H, and Dominique G Homberger. 2009. "Structure and Functions of Keratin Proteins in Simple, Stratified, Keratinized and Cornified Epithelia." *J. Anat* 214: 516–59. https://doi.org/10.1111/j.1469-7580.2009.01066.x.

Brown, Sam P, Helen E Blackwell, and Brian K Hammer. 2018. "GA.s 6th Conference on Cell-Cell Communication in Bacteria." *Jb.Asm.Org 1 Journal of Bacteriology* 200: 291–309. https://doi.org/10.1128/JB.

Cadena-Nava, Ruben D, Mauricio Comas-Garcia, Rees F Garmann, A L N Rao, Charles M Knobler, and William M Gelbart. 2012. "Self-Assembly of Viral Capsid Protein and RNA Molecules of Different Sizes: Requirement for a Specific High Protein/RNA Mass Ratio." *Journal of Virology* 86 (6): 3318–26. https://doi.org/10.1128/JVI.06566-11.

Calvaresi, Matteo, Leopold Eckhart, and Lorenzo Alibardi. 2016. "The Molecular Organization of the Beta-Sheet Region in Corneous Beta-Proteins (Beta-Keratins) of Sauropsids Explains Its Stability and Polymerization into Filaments." *Journal of Structural Biology* 194 (3): 282–91. https://doi.org/10.1016/j.jsb.2016.03.004.

Carlier, Marie France, and Dominique Pantaloni. 1997. "Control of Actin Dynamics in Cell Motility." *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.1997.1062.

Carvajal-Vallejos, P, R Pallissé, H D Mootz, and S R Schmidt. 2012. "Unprecedented Rates and Efficiencies Revealed for New Natural Split Inteins from Metagenomic Sources." *J Biol Chem* 287 (34): 28686–96. https://doi.org/10.1074/jbc.M112.372680.

Chen, Xiaoying, Jennica L. Zaro, and Wei-Chiang Shen. 2013. "Fusion Protein Linkers: Property, Design and Functionality." *Advanced Drug Delivery Reviews* 65 (10): 1357–69. https://doi.org/10.1016/j.addr.2012.09.039.

Chong, S, F B Mersha, D G Comb, M E Scott, D Landry, L M Vence, F B Perler, et al. 1997. "Single-Column Purification of Free Recombinant Proteins Using a Self-Cleavable Affinity Tag Derived from a Protein Splicing Element." *Gene* 192 (2): 271–81. http://www.ncbi.nlm.nih.gov/pubmed/9224900.

Chowdhury, C, S Chun, A Pang, M R Sawaya, S Sinha, T O Yeates, and T A Bobik. 2015. "Selective Molecular Transport through the Protein Shell of a Bacterial Microcompartment Organelle." *Proc Natl Acad Sci U S A* 112 (10): 2990–95. https://doi.org/10.1073/pnas.1423672112.

Chowdhury, Chiranjit, Sharmistha Sinha, Sunny Chun, Todd O Yeates, and Thomas A Bobik. 2014. "Diverse Bacterial Microcompartment Organelles." *Microbiology and Molecular Biology Reviews* 78 (3): 438–68. https://doi.org/10.1128/MMBR.00009-14.

Ciani, B, S Bjelic, S Honnappa, H Jawhari, R Jaussi, A Payapilly, T Jowitt, M O Steinmetz, and R A Kammerer. 2010. "Molecular Basis of Coiled-Coil Oligomerization-State Specificity." *Proc Natl Acad Sci U S A* 107 (46): 19850–55. https://doi.org/10.1073/pnas.1008502107.

Collaborative Computational Project, Number 4. 1994. "The CCP4 Suite: Programs for Protein Crystallography." *Acta Crystallographica Section D Biological Crystallography* 50 (5): 760–63. https://doi.org/10.1107/S0907444994003112.

Community, Synthetic Biology. 2003. "Synthetic Biology: Applications and Dimensions." Synthetic Biology Project. http://syntheticbiology.org/Applications.html.

Cortajarena, A. L., T. Kajander, W. Pan, M. J. Cocco, and L. Regan. 2004. "Protein Design to Understand Peptide Ligand Recognition by Tetratricopeptide Repeat Proteins." *Protein Engineering Design and Selection* 17 (4): 399–409. https://doi.org/10.1093/protein/gzh047.

Cortajarena, Aitziber L, Jimin Wang, and Lynne Regan. 2010. "Crystal Structure of a Designed Tetratricopeptide Repeat Module in Complex with Its Peptide Ligand." *FEBS Journal* 277 (4): 1058–66. https://doi.org/10.1111/j.1742-4658.2009.07549.x.

D'Andrea, L D, and L Regan. 2003. "TPR Proteins: The Versatile Helix." *Trends Biochem Sci* 28 (12): 655–62. https://doi.org/10.1016/j.tibs.2003.10.007.

Dai, Bin, Cameron J Sargent, Xinrui Gui, Cong Liu, and Fuzhong Zhang. 2019. "Fibril Self-Assembly of Amyloid–Spider Silk Block Polypeptides." *Biomacromolecules*, acs.biomac.9b00218. https://doi.org/10.1021/acs.biomac.9b00218.

Dassa, Bareket, Nir London, Barry L Stoddard, Ora Schueler-Furman, and Shmuel Pietrokovski. 2009. "Fractured Genes: A Novel Genomic Arrangement Involving New Split Inteins and a New Homing Endonuclease Family." *Nucleic Acids Res* 37 (8): 2560–73. https://doi.org/10.1093/nar/gkp095.

Debelle, Laurent, and Alain J.P. Alix. 2002. “The Structures of Elastins and Their Function.” *Biochimie* 81 (10): 981–94. https://pdf.sciencedirectassets.com/272115/1-s2.0-S0300908400X00363/1-s2.0-S0300908499002217/main.pdf?x-amz-security-token=AgoJb3JpZ2luX2VjEML%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJHMEUCIQD6FKBHXaKGMP2QE9qebiLsos9xceGRLut%2FZhR9PpIgNgIgVmyqg2K9Mn.

Dominguez, Roberto, and Kenneth C Holmes. 2011. “Actin Structure and Function.” *Annual Review of Biophysics* 40: 169–86. https://doi.org/10.1146/annurev-biophys-042910-155359.

Edwardson, Thomas G. W., Takahiro Mori, and Donald Hilvert. 2018. “Rational Engineering of a Designed Protein Cage for SiRNA Delivery.” *Journal of the American Chemical Society* 140 (33): 10439–42. https://doi.org/10.1021/jacs.8b06442.

Egloff, P, M Hillenbrand, C Klenk, A Batyuk, P Heine, S Balada, K M Schlinkmann, D J Scott, M Schütz, and A Plückthun. 2014. “Structure of Signaling-Competent Neurotensin Receptor 1 Obtained by Directed Evolution in Escherichia Coli.” *Proc Natl Acad Sci U S A* 111 (6): E655-62. https://doi.org/10.1073/pnas.1317903111.

Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. “Features and Development of *Coot*.” *Acta Crystallographica Section D Biological Crystallography* 66 (4): 486–501. https://doi.org/10.1107/S0907444910007493.

Fairhead, M, G Veggiani, M Lever, J Yan, D Mesner, C V Robinson, O Dushek, P A van der Merwe, and M Howarth. 2014. “SpyAvidin Hubs Enable Precise and Ultrastable Orthogonal Nanoassembly.” *J Am Chem Soc* 136 (35): 12355–63. https://doi.org/10.1021/ja505584f.

Fallas, Jorge A., George Ueda, William Sheffler, Vanessa Nguyen, Dan E. McNamara, Banumathi Sankaran, Jose Henrique Pereira, et al. 2017. “Computational Design of Self-Assembling Cyclic Protein Homo-Oligomers.” *Nature Chemistry* 9 (4): 353–60. https://doi.org/10.1038/nchem.2673.

Figueroa, M, N Oliveira, A Lejeune, K W Kaufmann, B M Dorr, A Matagne, J A Martial, J Meiler, and C Van de Weerdt. 2013. "Octarellin VI: Using Rosetta to Design a Putative Artificial (β/α)8 Protein." *PLoS One* 8 (8): e71858. https://doi.org/10.1371/journal.pone.0071858.

Finn, R D, A Bateman, J Clements, P Coggill, R Y Eberhardt, S R Eddy, A Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Res* 42 (Database issue): D222-30. https://doi.org/10.1093/nar/gkt1223.

Flenniken, M L, M Uchida, L O Liepold, S Kang, M J Young, and T Douglas. 2009. "A Library of Protein Cage Architectures as Nanomaterials." *Curr Top Microbiol Immunol* 327: 71–93. http://www.ncbi.nlm.nih.gov/pubmed/19198571.

Fletcher, J M, R L Harniman, F R Barnes, A L Boyle, A Collins, J Mantell, T H Sharp, et al. 2013. "Self-Assembling Cages from Coiled-Coil Peptide Modules." *Science* 340 (6132): 595–99. https://doi.org/10.1126/science.1233936.

Frangioni, J.V., and B.G. Neel. 1993. "Solubilization and Purification of Enzymatically Active Glutathione S-Transferase (PGEX) Fusion Proteins." *Analytical Biochemistry* 210 (1): 179–87. https://doi.org/10.1006/abio.1993.1170.

Franke, D., M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, et al. 2017. "*ATSAS 2.8*: A Comprehensive Data Analysis Suite for Small-Angle Scattering from Macromolecular Solutions." *Journal of Applied Crystallography* 50 (4): 1212–25. https://doi.org/10.1107/S1600576717007786.

Franke, Daniel, and Dmitri I Svergun. 2009. "DAMMIF, a Program for Rapid Ab-Initio Shape Determination in Small-Angle Scattering." *Journal of Applied Crystallography* 42 (Pt 2): 342–46. https://doi.org/10.1107/S0021889809000338.

Gradišar, Helena, Sabina Božič, Tibor Doles, Damjan Vengust, Iva Hafner-Bratkovič, Alenka Mertelj, Ben Webb, Andrej Šali, Sandi Klavžar, and Roman Jerala. 2013. "Design of a Single-Chain Polypeptide Tetrahedron Assembled from Coiled-Coil Segments." *Nature Chemical Biology* 9 (6): 362–66. https://doi.org/10.1038/nchembio.1248.

Green, E M, J C Mansfield, J S Bell, and C P Winlove. 2014. "The Structure and Micromechanics of Elastic Tissue." *Interface Focus* 4. https://doi.org/10.1098/rsfs.2013.0058.

Grove, Tijana Z., Jason Forster, Genaro Pimienta, Eric Dufresne, and Lynne Regan. 2012. “A Modular Approach to the Design of Protein-Based Smart Gels.” *Biopolymers* 97 (7): 508–17. https://doi.org/10.1002/bip.22033.

Grove, Tijana Z, Chinedum O Osuji, Jason D Forster, Eric R Dufresne, and Lynne Regan. 2010. “Stimuli-Responsive Smart Gels Realized via Modular Protein Design.” *Journal of the American Chemical Society* 132 (40): 14024–26. https://doi.org/10.1021/ja106619w.

Grove, Tijana Z, Lynne Regan, and Aitziber L Cortajarena. 2013. “Nanostructured Functional Films from Engineered Repeat Proteins.” *Journal of The Royal Society Interface* 10 (83): 20130051. https://doi.org/10.1098/rsif.2013.0051.

Harbury, P B, J J Plecs, B Tidor, T Alber, and P S Kim. 1998. “High-Resolution Protein Design with Backbone Freedom.” *Science* 282 (5393): 1462–67. http://www.ncbi.nlm.nih.gov/pubmed/9822371.

Harvey, Joseph A. 2016. “The Use of Native Chemical Ligation for Controllable Assembly of Protein-Based Nanostructures.”

Harvey, Joseph A., Laura S. Itzhaki, and Ewan R. G. Main. 2018. “Programmed Protein Self-Assembly Driven by Genetically Encoded Intein-Mediated Native Chemical Ligation.” *ACS Synthetic Biology* 7 (4): 1067–74. https://doi.org/10.1021/acssynbio.7b00447.

Hengen, Paul N. 1995. “Purification of His-Tag Fusion Proteins from Escherichia Coli.” *Trends in Biochemical Sciences* 20 (7): 285–86. https://doi.org/10.1016/S0968-0004(00)89045-3.

Hsia, Yang, Jacob B Bale, Shane Gonen, Dan Shi, William Sheffler, Kimberly K Fong, Una Nattermann, et al. 2016. “Design of a Hyperstable 60-Subunit Protein Icosahedron.” *Nature* 535. https://doi.org/10.1038/nature18010.

Kattula, Sravya, James R Byrnes, and Alisa S Wolberg. 2017. “Fibrinogen and Fibrin in Hemostasis and Thrombosis.” *Arteriosclerosis, Thrombosis, and Vascular Biology* 37 (3). https://doi.org/10.1161/atvbaha.117.308564.

Kerfeld, Cheryl A, and Onur Erbilgin. 2015. “Bacterial Microcompartments and the Modular Construction of Microbial Metabolism.” *Trends in Microbiology* 23 (1): 22–34. https://doi.org/10.1016/j.tim.2014.10.003.

Kim, Y E, Y N Kim, J A Kim, H M Kim, and Y Jung. 2015. “Green Fluorescent Protein Nanopolygons as Monodisperse Supramolecular Assemblies of Functional Proteins with Defined Valency.” *Nat Commun* 6: 7134. https://doi.org/10.1038/ncomms8134.

King, N P, J B Bale, W Sheffler, D E McNamara, S Gonen, T Gonen, T O Yeates, and D Baker. 2014. “Accurate Design of Co-Assembling Multi-Component Protein Nanomaterials.” *Nature* 510 (7503): 103–8. https://doi.org/10.1038/nature13404.

King, N P, W Sheffler, M R Sawaya, B S Vollmar, J P Sumida, I André, T Gonen, T O Yeates, and D Baker. 2012. “Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy.” *Science* 336 (6085): 1171–74. https://doi.org/10.1126/science.1219364.

Kohl, A, H K Binz, P Forrer, M T Stumpp, A Plückthun, and M G Grütter. 2003. “Designed to Be Stable: Crystal Structure of a Consensus Ankyrin Repeat Protein.” *Proc Natl Acad Sci U S A* 100 (4): 1700–1705. https://doi.org/10.1073/pnas.0337680100.

Kreplak, L, J Doucet, P Dumas, and F Briki. 2004. “New Aspects of the Alpha-Helix to Beta-Sheet Transition in Stretched Hard Alpha-Keratin Fibers.” *Biophysical Journal* 87 (1): 640–47. https://doi.org/10.1529/biophysj.103.036749.

Kudryashov, Dmitri S, and Emil Reisler. 2013. “ATP and ADP Actin States.” *Biopolymers* 99 (4): 245–56. https://doi.org/10.1002/bip.22155.

Lai, Y T, D Cascio, and T O Yeates. 2012. “Structure of a 16-Nm Cage Designed by Using Protein Oligomers.” *Science* 336 (6085): 1129. https://doi.org/10.1126/science.1219351.

Lawson, David M., Peter J. Artymiuk, Stephen J. Yewdall, John M. A. Smith, J. Craig Livingstone, Amyra Treffry, Alessandra Luzzago, et al. 1991. “Solving the Structure of Human H Ferritin by Genetically Engineering Intermolecular Crystal Contacts.” *Nature* 349 (6309): 541–44. https://doi.org/10.1038/349541a0.

Lee, J, S I Blaber, V K Dubey, and M Blaber. 2011. “A Polypeptide ‘Building Block’ for the β-Trefoil Fold Identified by ‘Top-down Symmetric Deconstruction.’” *J Mol Biol* 407 (5): 744–63. https://doi.org/10.1016/j.jmb.2011.02.002.

Lee, Jung-Lim, Cheon-Seok Park, and Hae-Yeong Kim. 2007. “Functional Assembly of Recombinant Human Ferritin Subunits in Pichia Pastoris.” *Journal of Microbiology and Biotechnology* 17 (10): 1695–99. http://www.ncbi.nlm.nih.gov/pubmed/18156787.

Ljubetič, A., I. Drobnak, H. Gradišar, and R. Jerala. 2016. “Designing the Structure and Folding Pathway of Modular Topological Bionanostructures.” *Chemical Communications* 52 (30): 5220–29. https://doi.org/10.1039/c6cc00421k.

Ljubetič, Ajasja, Fabio Lapenta, Helena Gradišar, Igor Drobnak, Jana Aupič, Žiga Strmšek, Duško Lainšček, et al. 2017. “Design of Coiled-Coil Protein-Origami Cages That Self-Assemble in Vitro and in Vivo.” *Nature Biotechnology* 35 (11): 1094. https://doi.org/10.1038/nbt.3994.

Main, E R, A R Lowe, S G Mochrie, S E Jackson, and L Regan. 2005. “A Recurring Theme in Protein Engineering: The Design, Stability and Folding of Repeat Proteins.” *Curr Opin Struct Biol* 15 (4): 464–71. https://doi.org/10.1016/j.sbi.2005.07.003.

Main, E R, K Stott, S E Jackson, and L Regan. 2005. “Local and Long-Range Stability in Tandemly Arrayed Tetratricopeptide Repeats.” *Proc Natl Acad Sci U S A* 102 (16): 5721–26. https://doi.org/10.1073/pnas.0404530102.

Main, E R, Y Xiong, M J Cocco, L D’Andrea, and L Regan. 2003. “Design of Stable Alpha-Helical Arrays from an Idealized TPR Motif.” *Structure* 11 (5): 497–508. http://www.ncbi.nlm.nih.gov/pubmed/12737816.

Markham, Nicholas R., and Michael Zuker. 2008. “UNAFold.” In , 3–31. Humana Press. https://doi.org/10.1007/978-1-60327-429-6_1.

Martin, D D, M Q Xu, and T C Evans. 2001. “Characterization of a Naturally Occurring Trans-Splicing Intein from Synechocystis Sp. PCC6803.” *Biochemistry* 40 (5): 1393–1402. http://www.ncbi.nlm.nih.gov/pubmed/11170467.

Mateu, Mauricio G. 2013. “Assembly, Stability and Dynamics of Virus Capsids.” *Archives of Biochemistry and Biophysics* 531: 65–79. https://doi.org/10.1016/j.abb.2012.10.015.

McKittrick, J., P. Y. Chen, S G Bodde, W Yang, E E Novitskaya, and M A Meyers. 2012. “The Structure, Functions, and Mechanical Properties of Keratin.” *JOM* 64 (4): 449–68. https://doi.org/10.1007/s11837-012-0302-8.

Mosavi, L K, D L Minor, and Z Y Peng. 2002. “Consensus-Derived Structural Determinants of the Ankyrin Repeat Motif.” *Proc Natl Acad Sci U S A* 99 (25): 16029–34. https://doi.org/10.1073/pnas.252537899.

Motojima, Fumihiro. 2015. “How Do Chaperonins Fold Protein?” *BIOPHYSICS* 11: 93–102. https://doi.org/10.2142/biophysics.11.93.

Pang, A, S Frank, I Brown, M J Warren, and R W Pickersgill. 2014. “Structural Insights into Higher Order Assembly and Function of the Bacterial Microcompartment Protein PduA.” *J Biol Chem* 289 (32): 22377–84. https://doi.org/10.1074/jbc.M114.569285.

Parsons, J B, S Frank, D Bhella, M Liang, M B Prentice, D P Mulvihill, and M J Warren. 2010. “Synthesis of Empty Bacterial Microcompartments, Directed Organelle Protein Incorporation, and Evidence of Filament-Associated Organelle Movement.” *Mol Cell* 38 (2): 305–15. https://doi.org/10.1016/j.molcel.2010.04.008.

Perlmutter, Jason D, and Michael F Hagan. 2015. “Mechanisms of Virus Assembly.” *Annual Review of Physical Chemistry* 66 (April): 217–39. https://doi.org/10.1146/annurev-physchem-040214-121637.

Petoukhov, Maxim V., Peter V. Konarev, Alexey G. Kikhney, and Dmitri I. Svergun. 2007. “*ATSAS 2.1* – towards Automated and Web-Supported Small-Angle Scattering Data Analysis.” *Journal of Applied Crystallography* 40 (s1): s223–28. https://doi.org/10.1107/S0021889807002853.

Phillips, J J, Y Javadi, C Millership, and E R Main. 2012. “Modulation of the Multistate Folding of Designed TPR Proteins through Intrinsic and Extrinsic Factors.” *Protein Sci* 21 (3): 327–38. https://doi.org/10.1002/pro.2018.

Phillips, Jonathan J., Charlotte Millership, and Ewan R. G. Main. 2012. “Fibrous Nanostructures from the Self-Assembly of Designed Repeat Protein Modules.” *Angew Chem Int Ed Engl* 51 (52): 13132–35. https://doi.org/10.1002/anie.201203795.

Pollard, Thomas D, Laurent Blanchoin, and R. Dyche Mullins. 2002. "Molecular Mechanisms Controlling Actin Filament Dynamics in Nonmuscle Cells." *Annual Review of Biophysics and Biomolecular Structure* 29 (1): 545–76. https://doi.org/10.1146/annurev.biophys.29.1.545.

Rath, Arianna, Mira Glibowicka, Vincent G. Nadeau, Gong Chen, and Charles M. Deber. 2009. "Detergent Binding Explains Anomalous SDS-PAGE Migration of Membrane Proteins." *Proceedings of the National Academy of Sciences* 106 (6): 1760–65. https://doi.org/10.1073/pnas.0813167106.

Rauscher, Sarah. 2017. "The Liquid Structure of Elastin Aggregates." *Biophysics and Structural Biology*, 1–21. https://doi.org/10.7554/eLife.26526.001.

Roos, W H, I L Ivanovska, A Evilevitch, and G J L Wuite. 2007. "Viral Capsids: Mechanical Characteristics, Genome Packaging and Delivery Mechanisms." *Cellular and Molecular Life Sciences*. https://doi.org/10.1007/s00018-007-6451-1.

Ross, James F, Angela Bridges, Jordan M Fletcher, Deborah Shoemark, Dominic Alibhai, Harriet E V Bray, Joseph L Beesley, et al. 2017. "Decorating Self-Assembled Peptide Cages with Proteins." https://doi.org/10.1021/acsnano.7b02368.

Ross, James F, Gemma C Wildsmith, Michael Johnson, Daniel L Hurdiss, Kristian Hollingsworth, Rebecca F Thompson, Majid Mosayebi, et al. 2019. "Directed Assembly of Homopentameric Cholera Toxin B-Subunit Proteins into Higher-Order Structures Using Coiled-Coil Appendages." https://doi.org/10.1021/jacs.8b11480.

SantaLucia, John, and Donald Hicks. 2004. "The Thermodynamics of DNA Structural Motifs." *Annual Review of Biophysics and Biomolecular Structure* 33 (1): 415–40. https://doi.org/10.1146/annurev.biophys.32.110601.141800.

Schlinkmann, K M, and A Plückthun. 2013. "Directed Evolution of G-Protein-Coupled Receptors for High Functional Expression and Detergent Stability." *Methods Enzymol* 520: 67–97. https://doi.org/10.1016/B978-0-12-391861-1.00004-6.

Sciore, Aaron, Min Su, Philipp Koldewey, Joseph D Eschweiler, Kelsey A Diffley, Brian M Linhares, Brandon T Ruotolo, James C A Bardwell, Georgios Skiniotis, and E Neil G Marsh. 2016. "Flexible, Symmetry-Directed Approach to Assembling Protein Cages." *Proceedings of the National Academy of Sciences of the United States of America* 113 (31): 8681–86. https://doi.org/10.1073/pnas.1606013113.

Scotter, A J, M Guo, M M Tomczak, M E Daley, R L Campbell, R J Oko, D A Bateman, A Chakrabartty, B D Sykes, and P L Davies. 2007. "Metal Ion-Dependent, Reversible, Protein Filament Formation by Designed Beta-Roll Polypeptides." *BMC Struct Biol* 7: 63. https://doi.org/10.1186/1472-6807-7-63.

Shah, Neel H, and Tom W Muir. 2014. "Inteins: Nature's Gift to Protein Chemists." *Chemical Science* 5 (1): 446–61. https://doi.org/10.1039/C3SC52951G.

Speltz, E B, A Nathan, and L Regan. 2015. "Design of Protein-Peptide Interaction Modules for Assembling Supramolecular Structures in Vivo and in Vitro." *ACS Chem Biol* 10 (9): 2108–15. https://doi.org/10.1021/acschembio.5b00415.

Spiess, Christoph, Anne S Meyer, Stefanie Reissmann, and Judith Frydman. 2004. "Mechanism of the Eukaryotic Chaperonin: Protein Folding in the Chamber of Secrets." *Trends Cell Biol* 14 (11): 598–604. https://doi.org/10.1016/j.tcb.2004.09.015.

Stranges, P B, and B Kuhlman. 2013. "A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds." *Protein Sci* 22 (1): 74–82. https://doi.org/10.1002/pro.2187.

Svergun, D., C. Barberato, and M. H. J. Koch. 1995. "*CRYSOL* – a Program to Evaluate X-Ray Solution Scattering of Biological Macromolecules from Atomic Coordinates." *Journal of Applied Crystallography* 28 (6): 768–73. https://doi.org/10.1107/S0021889895007047.

Svergun, D. I. 1992. "Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria." *Journal of Applied Crystallography* 25 (pt 4): 495–503. https://doi.org/10.1107/S0021889892001663.

Svergun, D I. 1999. “Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing.” *Biophysical Journal* 76 (6): 2879–86. https://doi.org/10.1016/S0006-3495(99)77443-6.

Svergun, D I, M V Petoukhov, and M H Koch. 2001. “Determination of Domain Structure of Proteins from X-Ray Solution Scattering.” *Biophysical Journal* 80 (6): 2946–53. https://doi.org/10.1016/S0006-3495(01)76260-1.

Uddin, Ismail, Stefanie Frank, Martin J. Warren, and Richard W. Pickersgill. 2018. “A Generic Self-Assembly Process in Microcompartments and Synthetic Protein Nanotubes.” *Small* 14 (19): 1704020. https://doi.org/10.1002/smll.201704020.

Undas, Anetta, and Robert A S Ariëns. 2011. “Brief Review Fibrin Clot Structure and Function A Role in the Pathophysiology of Arterial and Venous Thromboembolic Diseases.” https://doi.org/10.1161/ATVBAHA.111.230631.

Vagin, Alexei A., Roberto A. Steiner, Andrey A. Lebedev, Liz Potterton, Stuart McNicholas, Fei Long, and Garib N. Murshudov. 2004. “*REFMAC* 5 Dictionary: Organization of Prior Chemical Knowledge and Guidelines for Its Use.” *Acta Crystallographica Section D Biological Crystallography* 60 (12): 2184–95. https://doi.org/10.1107/S0907444904023510.

Voet, A R, H Noguchi, C Addy, D Simoncini, D Terada, S Unzai, S Y Park, K Y Zhang, and J R Tame. 2014. “Computational Design of a Self-Assembling Symmetrical β-Propeller Protein.” *Proc Natl Acad Sci U S A* 111 (42): 15102–7. https://doi.org/10.1073/pnas.1412768111.

Volkov, Vladimir V., and Dmitri I. Svergun. 2003. “Uniqueness of *Ab Initio* Shape Determination in Small-Angle Scattering.” *Journal of Applied Crystallography* 36 (3): 860–64. https://doi.org/10.1107/S0021889803000268.

Votteler, Jörg, Cassandra Ogohara, Sue Yi, Yang Hsia, Una Nattermann, David M Belnap, Neil P King, and Wesley I Sundquist. 2016. “Designed Proteins Induce the Formation of Nanocage-Containing Extracellular Vesicles.” *Nature* 540 (7632): 292–95. https://doi.org/10.1038/nature20607.

Wang, Bin, Wen Yang, Joanna McKittrick, and Marc André Meyers. 2016. "Keratin: Structure, Mechanical Properties, Occurrence in Biological Organisms, and Efforts at Bioinspiration." *Progress in Materials Science* 76 (March): 229–318. https://doi.org/10.1016/j.pmatsci.2015.06.001.

Watanabe, T, Y Ito, T Yamada, M Hashimoto, S Sekine, and H Tanaka. 1994. "The Roles of the C-Terminal Domain and Type III Domains of Chitinase A1 from Bacillus Circulans WL-12 in Chitin Degradation." *Journal of Bacteriology* 176 (15): 4465. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC196264/.

Weisel, John W, and Rustem I Litvinov. 2017. "Fibrin Formation, Structure and Properties." *Sub-Cellular Biochemistry* 82: 405–56. https://doi.org/10.1007/978-3-319-49674-0_13.

"What Is Synthetic Biology?" 2015. Synthetic Biology Project. http://www.synbioproject.org/topics/synbio101/definition/.

Wolberg, Alisa S, Robert A Campbell, and Alisa S Wolberg. 2008. "Thrombin Generation, Fibrin Clot Formation and Hemostasis." *Transfus Apher Sci* 38 (1): 15–23. https://doi.org/10.1016/j.transci.2007.12.005.

Woolfson, Derek N. 2014. "Assessing Cellular Response to Functionalized α-Helical Peptide Hydrogels." *Advanced Healthcare Materials* 3 (9): 1387–91. https://doi.org/10.1002/adhm.201400065.

Wright, James N. 2018. "Studies on the Self-Assembly of Geometrically Designed Protein Fusions Using Genetically Programmed Chemistry."

Wu, H, Z Hu, and X Q Liu. 1998. "Protein Trans-Splicing by a Split Intein Encoded in a Split DnaE Gene of Synechocystis Sp. PCC6803." *Proceedings of the National Academy of Sciences of the United States of America* 95 (16): 9226–31. http://www.ncbi.nlm.nih.gov/pubmed/9689062.

Yamagami, Motoya, Tomohisa Sawada, and Makoto Fujita. 2018. "Synthetic β-Barrel by Metal-Induced Folding and Assembly." *J. Am. Chem. Soc* 140: 15. https://doi.org/10.1021/jacs.8b04284.

Yin, Liang, Xiang Guo, Lu Liu, Yong Zhang, and Yan Feng. 2019. "Self-Assembled Multimeric-Enzyme Nanoreactor for Robust and Efficient Biocatalysis" 8: 5. https://doi.org/10.1021/acsbiomaterials.8b00279.

Zettler, J, V Schütz, and H D Mootz. 2009. “The Naturally Split Npu DnaE Intein Exhibits an Extraordinarily High Rate in the Protein Trans-Splicing Reaction.” *FEBS Lett* 583 (5): 909–14. https://doi.org/10.1016/j.febslet.2009.02.003.

Zhang, Yu, and Brendan P Orner. 2011. “Self-Assembly in the Ferritin Nano-Cage Protein Superfamily.” *International Journal of Molecular Sciences* 12 (8): 5406–21. https://doi.org/10.3390/ijms12085406.